APPLIED RESEARCH

SUMMARY

- Questions the assumption that usability methods are equally suitable for evaluating Web sites and Web applications
- Concludes that the decision to choose one variant of the think-aloud method over another should be based on practical considerations

Evaluation of an Informational Web Site: Three Variants of the Think-aloud Method Compared

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

INTRODUCTION

Usability methods such as concurrent or retrospective think-aloud protocols and constructive interaction were originally employed for physical tasks (Miyake 1982; Hoc and Leplat 1983). Gradually, they have also been discovered by usability experts to be a valuable tool for the testing of software applications such as multimedia authoring systems (O'Malley and colleagues 1991), computer software (Henderson and colleagues 1995), or e-mail applications (de Mul and van Oostendorp 1996). Since then, they have gained in popularity and have been applied to software applications as various as online library catalogs (Campbell 2001; Norlin and Winters 2002; Hall, de Jong, and Steehouder 2004), computer games (Höysniemi, Hämäläinen, and Turkki 2003), and groupware (Pinelle, Gutwin, and Greenberg 2003).

These days, the traditional usability methods are also—and increasingly—employed to evaluate informational Web sites (Sienot 1997; van Waes 2000; Benbunan-Fich 2001). Interestingly, this Web site evaluation is typically performed with the idea that standard usability methods will work exactly the same for testing Web sites as for testing software applications, revealing similar results regardless of the nature of the test object. As such, usability experts who evaluate a Web site often use the same methods that were originally designed for software applications. A good example of this straightforward adoption of methods concerns Nielsen's (1994) heuristics for evaluating software applications, which are often employed for Web sites without additional explanation or justification (see de Jong and van der Geest 2000).

However, it remains to be seen whether the traditional methods do indeed work for Web sites as they do for software applications. After all, there is a major difference between the two types of test objects. While both software applications and Web sites exist in various shapes, Web sites typically require more substantial reading, that is, processing of textual information, than software applications, which commonly require more physical actions, such as entering data or clicking on links or buttons. This difference results in a different degree of visibility of a user's dealings with a test object: a Web site evaluation, involving fewer physical actions, will often reveal fewer observable actions than an evaluation of a software application. As a consequence, usability experts wishing to detect user problems in a Web site are more dependent on the verbalizations of their participants than experts who are evaluating a software application.

Paradoxically, it is precisely the smaller number of observable actions that could make it harder for participants in a Web site evaluation to verbalize their thoughts. Recent studies have shown that participants in the evaluation of a Web application first and foremost tended to verbalize what they were physically doing at a particular moment (van den Haak, de Jong, and Schellens 2003, 2004). They would then move on from these verbalized actions to express further (more valuable) thoughts. The participants in a Web site evaluation, however, usually perform fewer visible actions, which gives them less opportunity to verbalize what they are doing and to use these verbalizations as a starting point for further thoughts. In other words, participants in a Web site evaluation typically receive fewer cues for verbalization than participants evaluating a Web application.

The different nature of (informational) Web sites and

Manuscript received 23 November 2005; revised 2 June 2006; accepted 6 June 2006.

(transactional) Web applications may have more specific consequences for the workings of the various usability methods. The concurrent think-aloud (CTA) method, for instance, is an evaluation tool that involves participants working with a test object and simultaneously verbalizing their thoughts. When employed for the evaluation of a Web application, this method has been known to cause reactivity among participants, who experienced an increased number of observable problems as a result of having to combine thinking aloud with performing mostly physical actions. Yet it is not unlikely that the same CTA method may cause a different kind of reactivity when used for the evaluation of an informational Web site. After all, participants having to combine thinking aloud with reading substantial bits of textual information may well process this information more carefully than had they been working silently (see Loxterman and colleagues 1994; Ummelen and Neutelings 2000). In that way, they may experience a decreased rather than increased number of observable problems.

A common variant of the CTA method is the retrospective think-aloud (RTA) method. This tool involves participants who work silently with a particular test object and afterwards comment on a recording of their performance. For this method, it would be interesting to see whether it is as effective for the testing of informational Web sites as it is for the testing of Web applications. Participants who verbalize on the basis of a recording of their task performance with a Web application will have sufficient visual cues to be able to recall their thoughts while working. However, if they are presented with a recording of an informational Web site evaluation, they may find fewer visual cues to support their verbalization task. This may not only lead to fewer or incomplete verbalizations (see Ericsson and Simon 1984) but may also make the participants more aware of and hence less at ease in the presence of the test facilitator.

A second variant of the CTA method is constructive interaction (called Team in our study), which involves two participants instead of one, who work together with a particular test object. This method has proven successful for the evaluation of software applications, but it remains to be seen how teams of two participants behave when confronted with large pieces of textual information. Reading is inherently an individual process in which it is hard to establish common ground, since it is not very likely that one or both participants will read out loud what they see on the screen. As such, it could be expected that a large part of the reading process will remain below the surface and that participants will only concentrate on those issues that require discussion.

In all, there is a real possibility that usability methods reveal different results depending on the test object to which they are applied. As such, we felt that it would be valuable to repeat, with a different test object, one of our previous studies (van den Haak, de Jong, and Schellens 2004), in which we investigated the workings of the above-mentioned three methods (CTA, RTA, and Team) for an online library catalog. The conclusion of that study was that while CTA, RTA, and Team revealed largely comparable results, give or take a few minor differences, the CTA method would seem most suitable for the testing of Web applications, as it offers practical advantages over the other two methods in terms of time and expense.

Since our previous study focused on a Web application, the current study will involve an informational Web site. To facilitate comparison of the current study to the previous one, we will address the same research questions that we posed before.

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

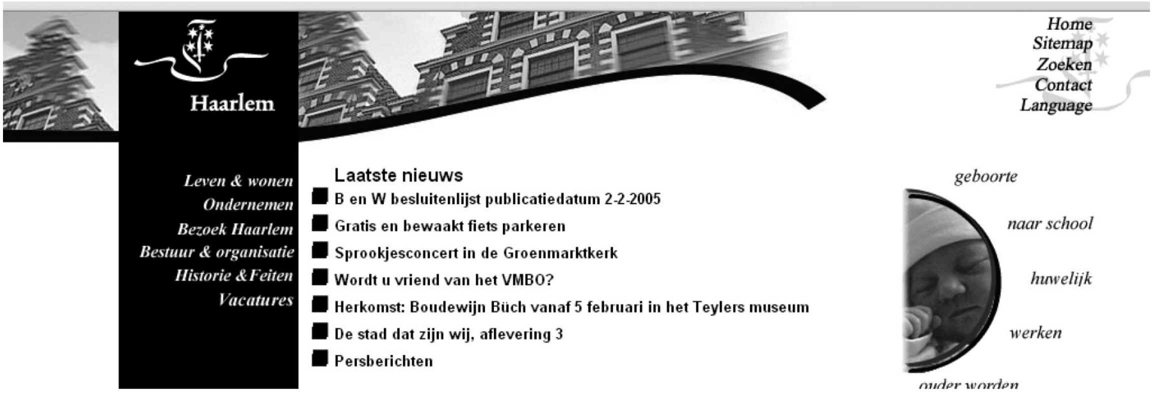
MATERIALS AND METHODS

Test object

Our test object was the municipal Web site of Haarlem (www.haarlem.nl), a city in the Netherlands that is home to about 150,000 people. The site is primarily intended for citizens of Haarlem, but it also offers information for those who plan to move there or simply wish to visit the city. Foreign visitors are also catered to, as some of the information on the site is available in English, German, and French.

As Figure 1 shows, the Web site has a simple layout, consisting of a home page with three columns. The column in the middle of the page offers the latest news related to Haarlem; this section is updated regularly. The left-hand column contains six main links that together cover all the standard information on the site. These links are “Living in Haarlem,” “Doing business in Haarlem,” “Visit Haarlem,” “Council and Organization,” “History and Facts,” and “Vacancies.” By clicking on these links, users are guided to the site’s various subsections. These subsections all contain a substantial piece of general information in the middle of the page as well as a number of sublinks that refer users to the specific information they are looking for. The “Living in Haarlem” link, for instance, contains sublinks on safety, taxes, health, and so forth, while the “Visit Haarlem” link has information on museums, restaurants, parking, shopping, and the like.

The right-hand column on the Haarlem Web site con-



Laatste nieuws

- B en W besluitenlijst publicatiedatum 2-2-2005
- Gratis en bewaakt fiets parkeren
- Sprookjesconcert in de Groenmarktkerk
- Wordt u vriend van het VMBO?
- Herkomst: Boudewijn Büch vanaf 5 februari in het Teylers museum
- De stad dat zijn wij, aflevering 3
- Persberichten

Figure 1. Home page of the Haarlem Web site.

tains largely the same information as the left-hand column, only here the information is presented more intuitively. The five links in this column form a half circle and represent the main stages of a life cycle, that is, birth, going to school, marriage, work, and getting older. The idea behind this cycle is that users often consult municipal sites when they or their children enter a stage in their lives that requires information exchange between them and the city. As such, the cycle offers users the opportunity to quickly find the information that they need based on the stage that they are in or about to enter at the time they access the site.

Participants

A total of 80 participants took part in our study. These participants, who had responded to printed and e-mail announcements and received a small financial compensation for their participation, were all students at the University of Twente, in Enschede. Most of them (55 students) were enrolled in Communication Studies; the other 25 took different courses. The average age of the students was 23, and the number of female and male participants was roughly equal (43 versus 37). At the time of the study, the students had spent on average three years at the university. While nearly all of them had occasionally worked with a municipal Web site before, none of them knew the Haarlem site. As such, the participants formed a suitable target group: they had experience with the kind of test object that was evaluated but not with the specific test object itself. All participants were evenly assigned to the three conditions in the study, with no difference regarding their demographic details.

Tasks

To evaluate the municipal Web site with the three usability test approaches, we formulated five main tasks, which were divided into 12 smaller subtasks. Each of the main tasks was introduced by means of a brief scenario description, which explained the context and provided details necessary to perform the tasks (marital status of the subject in the scenario, and so on). The performance of each task included not only a search component but also a substantial reading component. All tasks could be carried out independently from one another, to minimize the risk that participants would get stuck after one or two tasks. The entire set of tasks is presented in Figure 2.

Since informational Web sites are typically designed for a larger variety of users and users' needs than Web applications, whose design is often based on "precise specifications for a well-known group of customers" (de Marsico and Levialdi 2003), we made no attempt to evaluate the entire test object with the tasks that we formulated. Instead, we based our formulation of the tasks and scenarios on those parts of the site that contained information for people who were in the process of moving to Haarlem. In this way, given the fact that the participants in our study were residing in or near Enschede, and not in Haarlem, we could at least ensure the ecological validity of the tasks.

Questionnaires

Apart from carrying out the above tasks, participants were also required to fill in two questionnaires. The first questionnaire, which was given to the participants on entering the usability lab, contained questions about demographic details such as age, gender, and education. It also focused on the participants' experience in working with municipal Web sites.

The second questionnaire was designed to measure how the participants felt about their participation in the study. It focused on four aspects: (1) the participants' experiences on having to think aloud (concurrent or retrospectively) or work together, (2) the participants' estima-

1. Buying a house in Haarlem
You've found a part-time job in Haarlem and intend to combine your work with a part-time study. As you're still living in Enschede (a city at two hours distance from Haarlem), you start looking for a new place to live. After a short search, you find an apartment at €120,000 that you would like to buy.

1A. *Will the city council of Haarlem allow you to buy this apartment and live in it?*
1B. *Why (not)?*

2. Moving to Haarlem
You've found a suitable place to live in Haarlem. While moving there, you find an old fridge that's been left behind by the previous owners of the place. You realize that you cannot dump this fridge just anywhere, but you have no car to transport it to an official dump site.

2A. *Can you arrange for the fridge to be picked up from home?*
2B. *If so, are there any costs involved?*

3. Registering as a new inhabitant
Two weeks after you moved you realize that you have yet to register as a new inhabitant of Haarlem.

3A. *Can you register by post?*
3B. *If so, how does that work?*
3C. *Can you register online?*
3D. *If so, how does that work?*

4. Paying taxes
As a newly arrived house owner, you need to pay ownership taxes to the city of Haarlem. You have no idea, however, how high these taxes are and how you should pay them. All you know is that the estimated worth of your house amounts to €90,000.

4A. *What annual taxes do you need to pay?*
4B. *Can you pay them in instalments?*

5. Parking in Haarlem
While you don't have a car yourself, your partner, who is not residing in Haarlem, would like to be able to park close to your house. He/she asks you to look into the parking possibilities.

5A. *Can you obtain a parking license for your partner's car?*
5B. *Are there any other parking options that are free?*

Figure 2. Scenario-based tasks designed to evaluate the Haarlem Web site (translated from Dutch).

tion of their method of working on the five tasks (such as more versus less structured, faster versus slower than normal), (3) the participants' evaluation of the tasks that they performed (for example, "How satisfied are you with the tasks you performed?" and "How many tasks do you think you performed correctly?"), and (4) the participants' judgments about the presence of the facilitator and the recording equipment. For each of these four aspects, participants had to rate their experiences on a five-point scale 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 study after their task performance was completed. The participants in the retrospective think-aloud condition (RTA) received their second questionnaire in two parts: the first part, with questions relating to their method of working, was given to them once they completed their task performance; the second part, with questions on how they had experienced thinking aloud, was given to them as soon as the retrospective session was over.

Experimental procedure

Our study consisted of 60 sessions held separately in the same usability lab. Forty sessions involved the CTA and RTA participants; the remaining 20 sessions involved the Team participants, who participated in the study in teams of two. During each session, video recordings were made of the computer screen and the participants' voices, and a facilitator was present to observe the participants and take notes.

The experimental procedures of the three conditions were precisely the same as the procedures that we used in our previous experiment (van den Haak, de Jong, and Schellens 2004) to ensure a valid comparison between the present and previous study. For the sake of completeness, we include a description of these procedures below.

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 previous experience in working with municipal Web sites. After completing this questionnaire, the participant received the 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, told the participant to "think aloud while performing your tasks, and pretend that the facilitator is not there. Do not turn to her for assistance. If you fall silent for a while, the facilitator will remind you to keep talking aloud. Finally, remember that it is the municipal Web site, and not you, that is being tested." Once the participant had finished performing the tasks, he or she received the second questionnaire about the experience of participating in the test.

The experimental procedure in the Team condition was the following. As in the CTA condition, each participant 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, and the other next to it. They then received instructions that explicitly told them to work together: "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 test.

In the RTA condition, the experimental procedure started, once again, with the questionnaire on personal information and prior knowledge. As in the other two conditions, the participants were then given the tasks and oral instructions, but here they were instructed simply to 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 post-test 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 post-test questionnaire, with questions on how they had experienced thinking aloud retrospectively.

Processing of the data

Once the 60 sessions were completed, we made transcripts of all the Team, CTA, and RTA verbalizations, and charted all the participants' navigations through the municipal Web site. We then studied these navigations and other actions with a view to detecting usability problems that had arisen while the participants were using the Haarlem Web site. Our criterion for marking a particular situation as problematic was that it should deviate from what we had identified as the optimum working procedure for each task. In addition, we closely examined the transcripts and identified verbal indicators of problems experienced, such as expressions of doubt, task difficulty, incomprehensibility, or annoyance related to the use of the Web site.

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 based on how they had surfaced in the data: (1) through observation of the behavioral data, (2) through verbalization by the participant, or (3) through a combination of observation and verbalization. Finally, two independent coders divided all detected problems into nine specific problem types. These types (see Figure 3) are partly based on the categorization that we used in our previous studies (van den Haak, de Jong, and Schellens 2003, 2004) and partly based on the classification used by de Jong and Schellens (1998) for identifying problem types in brochures. The intercoder reliability was computed using Cohen's kappa. The overall kappa was .83, indicating a satisfactory level of inter-coder agreement.

Apart from the nine types of problems, participants also occasionally experienced technology problems, such as trouble with the network connection, the browser, or the computer used. In addition, a number of problems occurred that were not related to the site but to a participant's failure to read the task properly. One participant, for instance, wrongly assumed that he would be renting instead of buying an apartment; another participant acted on the assumption that he was living with someone rather than alone. Both the technology problems and the problems that were unrelated to the site were excluded from our analyses.

To determine the relevance of the problems detected, five independent experts rated each individual problem on a five-point Likert scale. Rating occurred twice, with the

1. **Comprehension**: Participant finds that the information on the site is not clear or applicable; he or she experiences syntax problems; he or she finds the choice of vocabulary problematic.

2. **Relevance**: Participant feels that certain information should not be included or should be reduced.

3. **Completeness**: Participant feels that information is missing or more elaboration is needed.

4. **Structure**: Participant finds that the order of information is problematic or that the structure is not clearly signaled.

5. **Formulation**: Participant does not appreciate a particular formulation.

6. **Graphic design**: Participant does not appreciate layout or illustrations.

7. **Correctness**: Participant detects a violation of syntax, spelling, or punctuation rules.

8. **Data entry**: Participant does not know how to enter data on the site.

9. **Visibility**: Participant fails to spot a particular link, button, or piece of information on the site.

Figure 3. Classification of problem types.

experts judging first the likelihood of the problem and then, on a separate occasion, its impact (presuming that the problem would be likely to recur) on the proper working of the Web site (see Nielsen 1994). The scores for the likelihood of the problems were multiplied by the scores for the problems' impact, resulting in a score for relevance. As these scores ranged from 1 to 25, we calculated their square roots as final scores for the relevance of problems.

To evaluate task performance in all three conditions, we used two indicators: the number of subtasks that were completed successfully and the time that was required to complete these tasks. We also investigated how the participants themselves felt about the tasks that they performed by analyzing their answers to the questions on task performance that were posed in the post-session questionnaire.

RESULTS

In this section we will first present the results of our analysis regarding the feedback (number and types of problems) collected with the three usability test approaches. We will then discuss the problems in terms of relevance, and the results with respect to task performance. We will conclude this section by describing how the participants experienced their participation in the study.

As we intended to investigate whether the usability approaches reveal different results when applied to a Web site rather than a Web application, the "Results" and "Discussion" sections of this article frequently refer to our previous study (van den Haak, de Jong, and Schellens 2004). In some instances, we will refer to two previous studies (van den Haak, de Jong, and Schellens 2003, 2004).

Number and types of problems detected

After analyzing the 60 recordings of sessions, we found a total of 119 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 among 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 verbalization, or (3) by a combination of observation and verbalization. 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) = 1.15, p = .323). This result is in line with the result of our previous experiment and thus reinforces the idea that each of the three methods is equally fruitful in terms of quantity of detected problems.

As for the way in which this output came about, there was only one significant difference, between the RTA and Team conditions. As is clear from Table 1, the participants in the RTA condition experienced significantly more observable problems than the participants in the Team condition (F(2,57) = 3.21, p < .05; Bonferroni post hoc analysis p < .05). This result might be caused by a different

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	1.8	1.9	1.1*	1.6	2.8*	2.8
Verbalized	2	2.1	2.4	2.8	3.4	2.5
Observed and verbalized	3.9	3	2.7	3.2	2.1	1.5
Total	7.7	3.7	6.2	5.9	8.3	3.8

◆ $^*p < 0.05$.

degree of attention required because of the large amount of information on the site: unlike the CTA participants (who had to think and read out loud and were therefore probably more focused on what they read) and the Team participants (who worked together and thus had two pairs of eyes available for reading), the RTA participants worked silently alone and may have done more skimming than reading of the site. In this way, they may have started clicking on links in the left- or right-hand column on the site before fully grasping the gist of each general piece of information in the middle of the site, thereby experiencing more observable problems.

Interestingly, the RTA condition did not reveal significantly more verbalized problems, unlike both our previous studies. This finding may be explained by the different nature of this study's test object: since the municipal Web site involved considerably more written information and fewer options for entering data than the library catalogs, the RTA participants spent most of their time skimming or reading text rather than entering search terms or clicking on buttons. However, as we indicated in the introduction to this article, processing text provides fewer cues than performing physical actions; thus, the RTA participants, while watching a recording of their task performance, received fewer stimuli to recall the problems they experienced. This situation prevented them from being more successful in verbalizing problems than the participants in either of the other two conditions.

As we predicted in the introduction to this article, the contribution of verbalizations to the CTA output was larger in the current study than in the previous study. While the percentage of purely verbalized problems was only slightly higher (26% versus 18%), the percentage of problems that were either detected or supported by verbalizations differed considerably (77% versus 43%).

To investigate the types of problems that were detected in the three conditions, we labeled all problems according to the problem types that we described above. Figure 4 shows a selection of problems as they occurred in the usability test approaches.

Table 2 shows the overall distribution of problem types in the CTA, Team, and RTA conditions. All participants clearly experienced most difficulties with the structure of the site. The results for the other problem types were quite similar across the three conditions too, with only one significant difference: the RTA condition revealed significantly more completeness problems than the CTA and Team conditions (F(2,57) = 5.164, $p < .05$; Bonferroni post hoc analysis $p < .05$).

A possible explanation for this difference lies in the fact that the RTA participants had additional time, while watching the recording of their performance, to reflect on the information they needed to perform their tasks and to determine whether the information was provided in the appropriate place. Alternatively, the difference could be explained in the same way as we explained the significant difference concerning the number of observable problems: since the RTA participants may have skimmed rather than read the text on the site, they may have overlooked certain information and as a result experienced more difficulty with the site that they attributed to a lack of completeness of information.

So far, we established two significant differences between the three usability test approaches as well as one noticeable lack of difference compared with the results of our previous study. The RTA condition revealed more observable problems than the Team condition, but not more verbalized problems than the CTA condition, as had been the case in our earlier experiment. In addition, the RTA participants experienced more completeness problems than the participants in the other two conditions. As we have shown, these results can all be explained by the text-oriented nature of the test object and the RTA participants' dealings with it.

We will now briefly consider the number of different problems detected in each condition (that is, the list of individual problems regardless of how many times they were detected) and the overlap between them. In the CTA condition, 64 different usability problems were detected; the Team condition revealed 74 different problems; in the RTA condition, 75 different problems came to light. Therefore, with respect to the range of individual problems detected, the Team and RTA methods were more profitable than the CTA method. This result is in line with our previous experiment, which also showed CTA as the least fruitful method.

1. **Comprehension**: The participant does not know the meaning of the label *schoon* (Dutch for “empty”) on a button on the “calculate your yearly taxes” page.

2. **Relevance**: There is too much information on the “parking in Haarlem” page.

3. **Completeness**: On the “tax” page, there is no information regarding paying taxes in installments.

4. **Structure**: The participant finds fault with the order of the links on the “coming and going” page.

5. **Formulation**: The participant objects to the use of the term *aangifte doen van een verhuizing* (Dutch for “reporting your move to Haarlem”), as it reminds him of reporting a crime to the police.

6. **Graphic design**: The participant is annoyed to find a different font size on the page containing information about parking in Haarlem.

7. **Correctness**: One of the two brackets is missing on an announcement put between brackets.

8. **Data entry**: The participant does not know how to enter a four-digit number in a box on the site.

9. **Visibility**: The navigation bar on the site is not clearly visible.

Figure 4. Examples of problem types as they occurred in the usability test approaches.

TABLE 2: TYPES OF PROBLEMS DETECTED PER PARTICIPANT IN THE CTA, TEAM, AND RTA CONDITION

	CTA		Team		RTA	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Comprehension	0.6	1	0.4	0.7	0.8	1.1
Relevance	0	0	0.1	0.2	0.2	0.4
Completeness	0.1*	0.2	0.1*	0.2	0.5*	0.8
Structure	5	2.6	4	4.2	5.5	3.2
Formulation	0	0	0	0	0.1	0.2
Graphic Design	0.9	0.8	0.8	1	0.7	0.7
Correctness	0.1	0.3	0.1	0.3	0.1	0.3
Data entry	0	0	0	0	0	0
Visibility	1	1	0.7	1	0.4	0.6

◆ ** RTA differs significantly from CTA and Team* ($p < 0.05$).

With respect to overlap in the three lists of usability problems, only 33 problems (28%) occurred in each of the three conditions. The overlap between two rather than three conditions was somewhat higher, ranging from 34% to 38%. These relatively low percentages indicate a substantial number of unique problems in each condition. This result is perhaps not very surprising given the volume (that is, the quality and quantity of pages) of the Web site that was tested. Nevertheless, if we take the frequency of the problems into account, the degree of overlap was considerable: problems that were detected in one condition by at least five participants were in 86% to 100% of the cases also detected by at least one participant in one of the other conditions. Thus, as in our previous experiment, each of the three methods could clearly predict the main output of the other two methods.

Relevance of the problems detected

As we mentioned above, five experts evaluated all 119 individual problems both in terms of likelihood and in terms of impact. Problems were rated on a Likert scale of 1 to 5 ("unlikely" to "highly likely" and "no impact" to "high impact"), and the two scores for each problem were multiplied. The square roots of these multiplied scores were taken as the final scores for relevance. These scores formed an adequately reliable scale (Cronbach's alpha = .72). With an average score of 3.07, the problems in the current study were rated as less relevant than the problems in the previous study, where the average relevance score was 3.43. An explanation for this discrepancy is that the problems in an informational Web site evaluation may be less univocally connected with task-related usability than the problems in the evaluation of a Web application. Participants have much more to see and find fault with, and may in the process of using the Web site comment on a wide variety of features.

With respect to the relevance of the problems detected in the three conditions, an analysis involving 95% confidence intervals showed that there were no significant differences: each of the methods proved equally useful in detecting relevant problems. There were also no significant differences with respect to the relevance of the problems that were unique to any of the three methods.

As for the manner in which the problems were detected, the CTA, RTA, and Team methods did not differ with respect to the relevance of problems detected through observation, verbalization, or both.

A final consideration involved the correlation between relevance and frequency of the usability problems. This correlation did not differ among the three conditions, and was generally low (r=0.18, p < .05). In other words, a problem that was frequently detected was not necessarily judged as more relevant than a problem that was less frequently detected. In our previous experiment, the correlation between frequency and relevance was considerably higher (r=0.46). The reason for this difference in correlation between the current and previous study is that in the current study, many of the conspicuous problems (such as a disproportionately big map or a change in font) had only a minor effect on the working of the Web site.

TABLE 3: TASK PERFORMANCE IN THE CTA, TEAM, AND RTA CONDITION

	CTA		Team		RTA	
	Mean	SD	Mean	SD	Mean	SD
Number of tasks completed successfully	10	2.3	11.3*	1	9.7*	1.7
Overall task completion time in minutes	25.1*	7.3	20.1*	3.7	22.2	6.5

◆ *$p < 0.05$.

Task performance

Two indicators were used to measure task performance in the three conditions: the number of subtasks that were completed successfully and the total amount of time required to complete the tasks. Table 3 shows the results of both indicators.

With regard to the completion of the 12 subtasks, we found one significant difference: the RTA participants completed fewer subtasks successfully (9.7) than the Team participants, who had a success rate of 11.3 subtasks (ANOVA, F(2,57) = 4.75, p < .05; Bonferroni post hoc analysis, p < .05). This difference is in line with the significantly larger number of observable problems in the RTA condition and may be explained in a similar manner: the RTA participants, in presumably skimming rather than reading the texts on the site, may have made more mistakes in completing their tasks.

The CTA participants performed their subtasks neither better nor worse than the participants in the other two conditions. This result corresponds to the findings of our previous study and suggests that the task performance of the participants in the CTA condition was not affected by reactivity, that is, by the double workload of having to think aloud while performing their tasks.

With regard to the overall task completion time, there was also one significant difference: the participants in the Team condition took less time (20.1 minutes) to complete their subtasks than the CTA participants (25.1 minutes) (ANOVA, F(2,57) = 3.53, p < .05; Bonferroni post hoc

TABLE 4: PARTICIPANTS' EVALUATION OF THE TASKS THAT THEY PERFORMED

	CTA		Team actor		Team co-actor		RTA	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
How satisfied/unsatisfied are you with the tasks performed?	2*	0.7	1.5	0.5	1.2*	0.5	2*	0.9
How easy/difficult did you think the tasks were?	2.6	0.8	2.1	1.1	2.3	1.2	2.8	1
How many tasks do you think you performed correctly?	3.9**	0.6	4.7**	0.5	4.6**	0.5	4.2	0.8

- *Note: Scores for "satisfaction" and "ease of task performance" are indicated on a five-point scale (1 = very satisfied, 5 = very unsatisfied, and so on).*
- ** Team co-actor differs significantly from CTA and RTA ($p < 0.05$).*
- *** CTA differs significantly from Team actor and Team co-actor ($p < 0.05$).*

analysis, p < .05). This result is interesting, as it conflicts with the findings of our previous study, in which the Team participants needed more time than the RTA participants to complete their tasks.

A possible explanation for this notable difference between the two studies can, once again, be found in the test objects that were used. Working with the online library catalog, as the participants in the previous study did, involved performing a substantial number of physical actions (entering search terms, selecting titles from a result list, and so on). These actions could often be carried out in more than one way, so the pairs of Team participants had plenty of opportunity to suggest possible ways of performing a task, hence taking more time to complete it. In the present study, however, working with the test object primarily involved reading and finding the answer to a particular question. Such actions are typically less open to alternative ways of performing them; after all, once the text containing the appropriate piece of information is found, there is no need to continue looking for it elsewhere. This fact, combined with the fact that the Team participants had two pairs of eyes instead of one to find relevant information, may have caused them to complete their tasks within a shorter period of time.

As we mentioned earlier on in this article, we also investigated how the participants themselves felt about the tasks they performed. To this purpose, the post-test questionnaire on participant experiences contained three questions on task performance, the first two of which were answered on a five-point scale: (1) How satisfied or unsatisfied are you with the tasks you performed? (2) How difficult or easy did you think the tasks were? (3) How many tasks do you think you performed correctly?

Since the participants in the Team condition were working in pairs, each with a different role (actor or observer) that may have affected their experiences, they will be treated as separate subgroups in the analyses of the questionnaire. Team actor will represent the actors (those working behind the computer); Team co-actor will represent the co-actors (those sitting next to the person working behind the computer).

The results of the questions on task performance are presented in Table 4. There were two significant differences between the conditions: the co-actors were more satisfied with their performance than the CTA and RTA participants (ANOVA, F(3,76) = 6.21, p < .01; Bonferroni post hoc analysis, p < .05), and both the actors and co-actors were more optimistic about the number of correctly performed tasks than the CTA participants (ANOVA, F(3,76) = 7.26, p < .01; Bonferroni post hoc analysis, p < .05). Both differences might be explained by the fact that the Team participants, in working together, felt more confident about their performance than the participants in the other two conditions.

Participant experiences

This section reports on the remaining questions in the questionnaire on participant experiences. These questions involved three aspects: (1) Experiences with having to think aloud (concurrently or retrospectively) or work together; (2) Method of working; (3) Presence of the facilitator and the recording equipment. As in the previous

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	2.8*	0.8	2.9*	0.8	3	0.6
More-less focused	2.7	0.7	3	0.9	2.8	0.6	2.4	0.8
More-less concentrated	2.8	0.4	3.2	0.7	3	0.5	2.9	0.7
More-less persevering	2.8	0.6	2.7	0.7	3.1	0.8	2.8	0.6
More-less successful	3	0.3	2.7	0.7	3	0.3	3	0.3
More-less pleasant	3.2	0.4	2.9	0.4	3	0.6	3.1	0.5
More-less accurate	2.8	0.6	3.2**	0.4	3.1	0.5	2.7**	0.7
More-less stressful	3.4	0.8	3	0.6	3.1	0.6	3.2	0.5

◆ *Note: Scores on a five-point scale (3 = no difference from usual).*
◆ ** CTA differs significantly from Team actor and Team co-actor ($p < 0.05$).*
◆ *** $p < 0.05$.*

section, which dealt with the questions on task performance, we will treat the Team actors and co-actors as separate subgroups.

To start with, all participants were asked how they had felt about having to think aloud (concurrently or retrospectively) or work 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 = .88). ANOVA testing and Bonferroni post hoc analyses showed that both the actors (mean score = 1.5) and the co-actors (mean score = 1.6) were significantly more positive about the Team method than the CTA participants (mean score = 2.8) and RTA participants (mean score = 3.1) about their respective methods (ANOVA, F(3,76) = 31, $p < .01$; Bonferroni post hoc analysis, $p < .05$). These results are the same as those of our previous study and thus reinforce the idea that constructive interaction is experienced more positively by participants than the concurrent or retrospective think-aloud methods.

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, and so on, they had worked than they would normally do. The results in Table 5 show 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 two of the eight items showed significant differences among the conditions. The first item involved the speed at which the participants performed their tasks. Compared with their normal working procedure, the CTA participants were less optimistic in their estimation of how fast they had worked than the actors and co-actors (ANOVA, F(3,76) = 4.3, $p < .05$; Bonferroni post hoc analysis, $p < .05$).

The second item that revealed a difference among the conditions concerned the self-reported accuracy with which the participants believed that they had performed their tasks. The RTA participants felt that they had been more accurate than the Team actors indicated that they had been (ANOVA, F(3,76) = 3.57; $p < .05$; Bonferroni post hoc analysis, $p < .05$). This finding is remarkable since we have shown above that the RTA participants experienced more observable problems than the Team participants as well as more completeness problems than the CTA and Team participants. Apparently, the RTA participants were more optimistic about the accuracy of their performance than they had reason to be.

The final part of the questionnaire included questions about the presence of the facilitator and the use of recording equipment. A first question, measured on a five-point scale, involved the degree of awareness of the facilitator who was present at the sessions. The participants indicated that they were somewhat but not fully aware of the facilitator's presence, with average scores ranging from 2 to 2.6. There was one significant difference among the conditions: the RTA participants (mean score = 2) were more aware of the facilitator being present than the Team actors (mean score = 2.5) and the Team co-actors (mean score = 2.6) (ANOVA, F(3,76) = 3.92, $p < .05$; Bonferroni post hoc analysis, $p < .05$). This difference can once again be explained by the nature of the test object: as the municipal Web site, consisting primarily of text to be read/skimmed, offered few visual stimuli for the RTA participants while watching their performance in retrospect, there were occasions when the participants could not recall what they had been doing at a particular moment and thus fell silent. On these occasions, the RTA participants may have been more aware of the test facilitator observing them.

Participants were then asked to indicate, once again on a five-point scale, to which degree they found it pleasant or unpleasant, natural or unnatural, and not disturbing or disturbing to have the facilitator present during the study. They were 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 = .86), 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 among the conditions. As the average scores of the participants ranged between 1.9 and 2.5, the participants clearly felt that they were not affected by the experimental setting. For the RTA participants this meant that even though they were more aware of the test facilitator, her presence did not particularly bother them.

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

DISCUSSION

As we pointed out in the introduction to this article, we were interested to learn whether the three variants of the think-aloud method that we previously used to test Web applications (online library catalogs) would reveal different results when applied to an informational Web site. The results of our current study showed that in some respects, there were indeed differences between the previous and current study. Regardless of the method used, there were two differences relating to the relevance of the problems detected. The output of the informational Web site evaluation was on average rated as less relevant than the output regarding the Web application, and the correlation between frequency and relevance of the problems detected was considerably lower for the informational Web site evaluation than for the evaluation of the Web application.

Both of these differences can be explained by the fact that informational Web sites typically have a more substantial volume than Web applications and that not all of this volume can be directly linked to the usability of the sites. A practical implication for usability experts is that they cannot simply assume that the problems detected by their participants will be equally relevant. As a result, the usability experts will have to rely, more so than in the evaluation of a Web application, on their own subsequent estimation of the severity of the problems detected with the informational Web site.

The workings of the individual methods in the evaluation of informational Web sites versus Web applications also reveal a number of differences caused by the nature of the test objects involved. The results of the Team condition showed that constructive interaction proved faster in the evaluation of an informational Web site than in the evaluation of a Web application. A further advantage of this method is that, as in the previous experiment, it received a more positive participant evaluation than the other two methods. There is, however, a potential drawback to using constructive interaction for informational Web site evaluation and that concerns the ecological validity of the method: it is questionable whether two people in a real-life situation would work together on a Web site such as the one we tested.

As for the RTA method, the participants in this condition experienced relatively more observable problems, completed fewer tasks successfully, and were more aware of the test facilitator—all due to the nature of the site, which presumably led them to skim rather than read and gave them fewer cues. As such, it would seem that the RTA method appears less suitable for the testing of informational Web sites than for the testing of Web applications. Nevertheless, RTA remains the most ecologically valid method of the three.

Contrary to the RTA method, the CTA method would seem more suitable for the testing of informational Web sites than for the testing of Web applications. Compared with our two previous experiments, the CTA participants in the current study detected fewer observable problems, which could suggest that they experienced less reactivity in working with the municipal Web site than in working with the online library catalogs. Moreover, as the role of verbalizations in revealing or supporting problem detections was larger in the informational Web site evaluation than in the evaluation of the Web application, we could argue that in the current study, the CTA method more clearly lives up to its name.

Despite these differences, there are still a number of important similarities between the previous and the current study. Regardless of whether they were used to evaluate a Web application or an informational Web site, the three methods did not differ with respect to the total number of problems detected. Moreover, the methods were equally successful in detecting relevant problems, and each method proved capable of predicting the main output of the other two methods. As such, we would have to conclude that the three methods are largely interchangeable—that is, they can be employed as effectively for the evaluation of informational Web sites as for the evaluation of Web applications.

On a practical note, the interchangeability of the methods means that the choice of one rather than another method should be dependent on usability experts' own priorities. If experts are primarily interested in finding usability problems, they would be advised to use the CTA method, since it is less expensive than constructive interaction and less time-consuming than RTA. However, experts who are also interested in a truthful representation of task performance would profit more from the RTA method, while those who feel that it is vital that their participants experience their usability test as being as pleasant as possible will be likely to opt for constructive interaction. TC

REFERENCES

Benbunan-Fich, R. 2001. Using protocol analysis to evaluate the usability of a commercial Web site. *Information and management* 39:151–163.

Campbell, N., ed. 2001. *Usability assessment of library-related Web sites: Methods and case studies.* Chicago, IL: LITA.

de Jong, M., and P. J. Schellens. 1998. Focus groups or individual interviews? A comparison of text evaluation approaches. *Technical communication* 45:77–88.

de Jong, M., and T. van der Geest. 2000. Characterizing Web heuristics. *Technical communication* 47:311–326.

de Marsico, M. Levialdi, and S. Levialdi. 2004. Evaluating Web sites: Exploiting user's expectations. *International journal of human-computer studies* 60:381–416.

de Mul, S., and H. van Oostendorp. 1996. Learning user interfaces by exploration. *Acta psychologica* 91:325–344.

Ericsson, K. A., and H. A. Simon. 1993. *Protocol analysis: Verbal reports as data.* Rev. ed. Cambridge, MA: MIT Press.

Hall, M., M. de Jong, and M. Steehouder. 1994. Cultural differences and usability evaluation: Individualistic and collectivistic participants compared. *Technical communication* 51:489–503.

Henderson, R. D., M. C. Smith, J. Podd, and H. Varela-Alvarez. 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 J. Leplat. 1983. Evaluation of different modalities of verbalisation in a sorting task. *International journal of man-machine studies* 18:283–306.

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

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

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

Nielsen, J., and R. L. Mack, eds. 1994. *Usability inspection methods.* New York, NY: Wiley.

Norlin, E., and C. M. I. Winters. 2002. *Usability testing for library Websites: A hands-on guide.* Chicago, IL: American Library Association.

O'Malley, C., M. Baker, and M. Elsom-Cook. 1991. The design and evaluation of a multimedia authoring system. *Computers & education* 17:49–60.

Pinelle, D., C. Gutwin, and S. Greenberg. 2003. Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM transactions on human-computer interaction* 10:281–311.

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.

Ummelen, N., and R. Neutelings. 2000. Measuring reading behavior in policy documents: A comparison of two instruments. *IEEE transactions on professional communication* 43:292–301.

van den Haak, M. J., M. D. T. de Jong, and P. J. Schellens. 2003. Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & information technology* 22:339–351.

———. 2004. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with computers* 16:1153–1170.

van Waes, L. 2000. Thinking aloud as a method for testing the usability of Web sites: The influence of task variation on the evaluation of hypertext. *IEEE transactions on professional communication* 43:279–291.

MAAIKE J. VAN DEN HAAK is a part-time PhD candidate at the University of Twente, the Netherlands. Her PhD research focuses on the merits and drawbacks of variants of the think-aloud method as an evaluation tool for instructive communication. Apart from her position at the University of Twente, she is also a part-time teacher of English and translation at the Vrije Universiteit in Amsterdam. Contact mj.vanden.haak@let.vu.nl.

MENNO DE JONG is an associate professor of communication studies at the University of Twente, the Netherlands. His main research interest concerns the methodology of applied communication research. He has published many articles about document and Web site evaluation and usability testing, and is currently working on an additional research line on applied research methods in organizational and corporate communication. Contact M.D.T.deJong@utwente.nl.

PETER JAN SCHELLENS is a professor of verbal communication at the faculty of Arts (Centre for Language Studies) of the Radboud University Nijmegen, the Netherlands. His research interests include document design, text- and Web-evaluation, and argumentation theory. Contact P.Schellens@let.ru.nl.