Government Information Quarterly 26 (2009) 193–202

Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Evaluating municipal websites: A methodological comparison of three think-aloud variants

Maaike J. van den Haak *, Menno D.T. de Jong, Peter Jan Schellens

Vrije Universiteit, Faculty of Arts, Netherlands
University of Twente, Faculty of Behavioural Sciences, Netherlands
Radboud University Nijmegen, Centre for Language Studies, Netherlands

ARTICLE INFO

Available online 29 October 2008

Keywords:
E-Government
Municipal websites
Usability testing
Think-aloud protocols

ABSTRACT

Usability methods have received relatively little methodological attention within the field of E-Government. This paper aims to address this gap by reporting on a usability test of the municipal website of Deventer (the Netherlands), carried out by means of three variants of the think-aloud method (concurrent/retrospective think-aloud protocols and constructive interaction). These three methods had proved successful in a previous evaluation of a different municipal website, yet we decided to replicate our study in order to investigate whether the three methods would reveal different results when applied to another municipal website with a different information architecture. The results of our study showed that, as in the previous municipal website evaluation, the three evaluation methods were largely comparable in terms of output. Nevertheless, we did find a number of differences between the present and previous municipal website evaluation regarding the workings of the three methods—differences that could be explained by the different information architectures of the municipal websites tested. This suggests that the three evaluation methods might indeed work differently depending on the nature of the website that is being evaluated, and calls for more research into the effect of task type on the validity of evaluation methods.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

Most municipalities and government institutions have their own space on the web, allowing their citizens to find information and, increasingly, to engage in all sorts of personalized E-Government services (Pieterson, Ebbers & Van Dijk, 2007). Citizens may, for instance, order copies of brochures, report changes in their address, or renew their vehicle registration, and the list of possibilities is likely to grow.

As the online activities of municipalities increased, so has the research output on this particular area (Heeks & Bailur, 2007). Studies have been published on municipal websites from countries as diverse as Norway (Halland & Saeth, 2004), Switzerland (Schedler & Summermatter, 2007), New Zealand (Cullen & Houghton, 2000), and Kenya (Kaaya, 2004). Some reports have even investigated the online activities of municipalities worldwide (Choudrie, Ghinea & Weerakkody, 2004; Holzer et al., 2006). The topics addressed in these studies vary widely from the accessibility of municipal websites (Potter, 2003; Shi, 2007; Jaeger, 2006) to textual content (Eschenfelder, 2004; Eschenfelder & Miller, 2007) to government–citizen interaction (Chadwick & May, 2003; Welch & Fulla, 2005; Griffin & Halpin, 2005).

Another major concern within the literature on E-Government is website usability. Numerous studies within E-Government journals, including GIQ, report on municipal website evaluations performed by means of usability methods like heuristics (Cullen & Houghton, 2000; Choudrie et al., 2004), scenario evaluation (Halland & Saeth, 2004; De Jong & Lentz, 2006b), interviews (Marcella, Baxter & Moore, 2003), and think-aloud protocols (Marcella et al., 2003; Jaeger, 2006). In describing the results of these website evaluations, however, most studies focus on the merits and drawbacks of the websites rather than on the working of the usability methods. As such, it seems that much is known about the ways in which municipal websites could be improved, but only little is known about the drawbacks and benefits of using a particular method for a particular municipal website. Since the validity of results revealed by usability methods is largely dependent on the validity of the methods themselves, more research into the exact working of usability methods within the field of E-Government seems highly desirable (see also Heeks and Bailur, 2007).

As a first step towards addressing this lack of attention for usability methodology within the E-Government area, we have recently evaluated a municipal website by means of three research methods: concurrent think-aloud protocols (CTA protocols), retrospective think-aloud protocols (RTA protocols), and constructive interaction (Van den Haak, De Jong & Schellens, 2007). The CTA protocols are perhaps the most common method of the three. They involve potential users who work with a particular test object while constantly verbalizing their

* Corresponding author. Vrije Universiteit, Faculty of Arts, Netherlands.
E-mail address: mj.vanden.haak@let.vu.nl (M.J. van den Haak).

0740-624X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.giq.2007.11.003

thoughts. The RTA protocols are a variant of the CTA protocols, involving participants who silently work with a particular test object and then verbalize their thoughts afterwards, often on the basis of a video recording of their performance. Constructive interaction, finally, is a method which involves two participants rather than one. They work together and verbalize their thoughts by interacting with each other. The practical value of the three methods is that researchers cannot only *observe* participants while working with a particular test object, but can also *listen* to them, primarily with a view to either uncovering people's mental processes or, in the case of usability evaluation, detecting user problems.

All three methods have long been accepted as useful research methods and have been applied in various fields including psychology (e.g. Taylor & Dionne, 2000), nursing (e.g. Funkesson, Anbäcken & Ek, 2007), and reading and writing research (Schellings, Aarnoutse & Van Leeuwe, 2006; Wong, 2005). Particularly the CTA and RTA protocols have been extensively discussed in research contributions, with Ericsson and Simon (1993) as standard theoretical framework. Within the context of usability testing, Nielsen (1993) is an often-cited practical handbook. Van Someren, Barnard and Sandberg (1994) also offer practical advice for the entire process of collecting and analyzing think-aloud protocols, as do Rubin (1994) and Dumas and Redish (1999) for the broader context of usability testing. Current research into the think-aloud methods has focused on, for instance, the effect of personality traits on people's ability to think aloud (Schneider & Reichl, 2006) and the notion of reactivity. This notion refers to the fact that when asked to perform tasks *and* think aloud at the same time, participants might perform these tasks differently and might experience difficulty in verbalizing, as a result of their combined cognitive workload being too high. As such, reactivity might affect the working of the concurrent think-aloud method. The extent to which this happens has been and continues to be a much investigated topic (Russo et al., 1989; Ericsson and Simon, 1993; Van den Haak, De Jong & Schellens, 2003; Van den Haak, De Jong & Schellens, 2004; Alavi, 2005).

The results of our study involving the three usability methods (Van den Haak et al., 2007) suggested that each of the methods was equally useful for evaluating municipal websites: the CTA protocols, RTA protocols, and constructive interaction were all comparable in terms of quantity and relevance of problems detected. Each of the methods was equally capable of detecting the main output of the other two methods. We did, however, find some differences between the three methods. The participants in the RTA method, for instance, experienced more observable problems and were less successful in completing their tasks than the participants in the CI method, while the CI participants performed their tasks faster than the CTA participants.

While these findings are interesting, we felt that it would be good to conduct a second municipal website evaluation using the same three usability methods but a different municipal website. The main reason for this replication of our previous study is that even within one country there are many municipal websites (De Jong and Lentz, 2006a) and these may vary greatly in terms of information architecture. The municipal website of our previous evaluation (Van den Haak et al., 2007), for instance, contains large pieces of information on every web page, and a list of links from which users, once they have read the information on the page, can make deliberate selections. Such an information architecture involves substantial

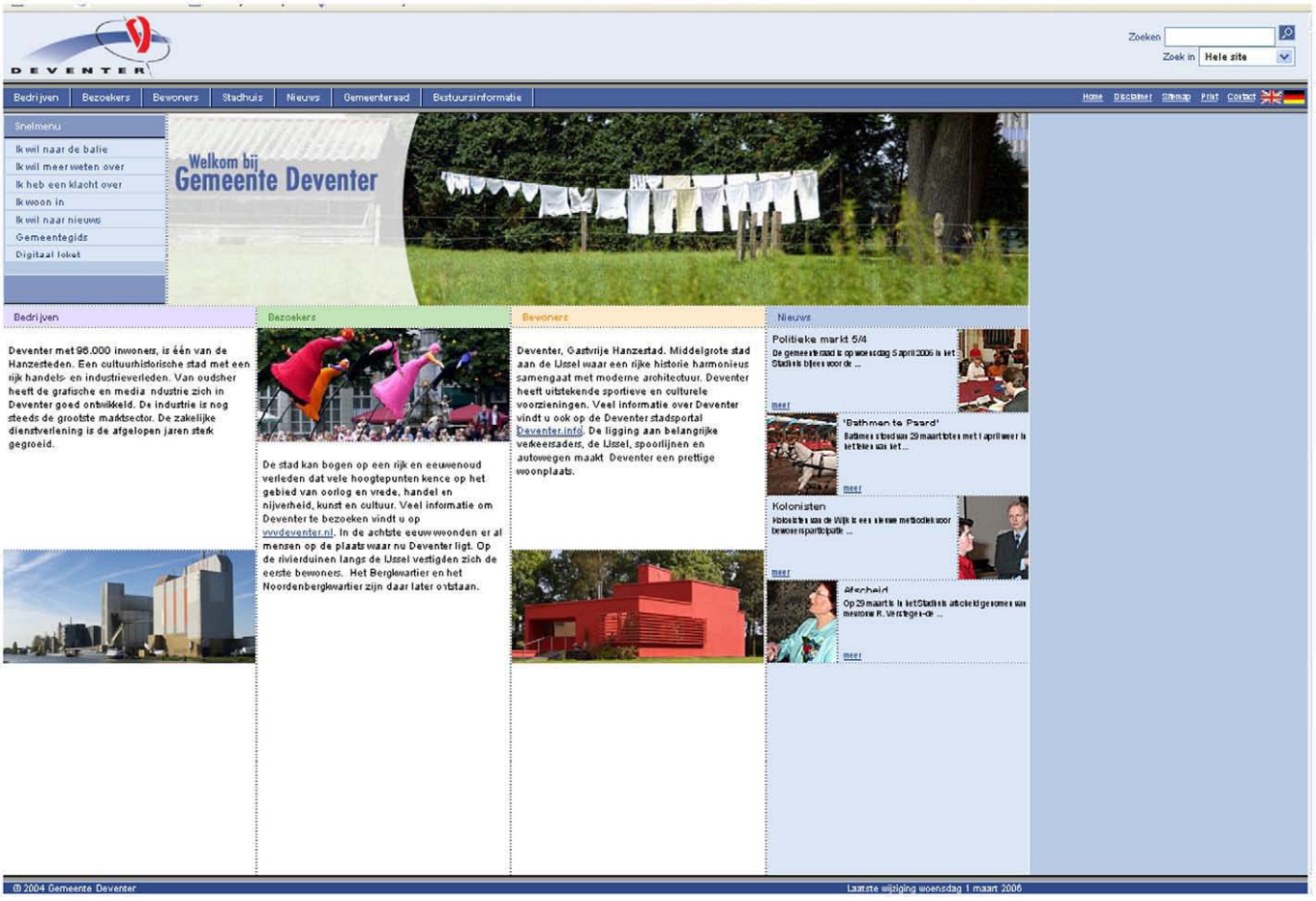

Fig. 1. Home page of the Deventer website.

reading and relatively little navigating, yet other municipal websites with different information architectures may involve different degrees of navigating and reading, or other types of tasks altogether (filling out web forms, etc.). These different task types could potentially affect the working of the three think-aloud methods. It is well imaginable, for instance, that participants working together in the CI method will find it more difficult to perform reading tasks than navigating tasks, since the former involve an inherently individual process, while the latter involve physical actions that are visible to both participants and can thus be discussed more easily. Likewise, when recorded on video or DVD, the physical nature of navigational tasks will probably be more informative for RTA participants to base their retrospective verbalizations on than tasks involving reading. This could mean that participants retrospectively verbalize more problems when they have just engaged in navigating than when they have just engaged in substantial reading. Finally, CTA participants might well experience more or less reactivity depending on whether they perform navigational or reading tasks. While reading, the fact that these participants have to think aloud might make them more attentive to the reading material at hand (see also Ummelen and Neutelings (2000) for some evidence to this claim), causing their performance to improve (positive reactivity). While navigating, on the other hand, participants have to think aloud and engage in physical activities, which could increase their cognitive workload and might result in a worse performance (see also Van den Haak et al., 2007).

It seems possible, then, that the results we found on the basis of our first municipal website evaluation (Van den Haak et al., 2007) will not apply when we employ the three usability methods to municipal websites involving different degrees of navigating and reading. As such, we have selected a second municipal website with a distinctly different information architecture from the previous one (see the section 'Test object' below for further details). With this website we performed an identical evaluation using the same three evaluation methods and research questions from our previous study:

- 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 problems detected?
- Do the three methods differ in terms of task performance?
- Do the three methods differ in terms of participant experiences?

In this way, we hoped to gain a more substantiated picture of how the three methods work when applied to municipal websites. The structure of this article is as follows. We will first discuss our research procedure, including details on participants, data collection, data processing, etc. We then describe the results of our present evaluation, and round off with a discussion section in which we compare the results of the present evaluation to those of the previous evaluation.

2. Method

2.1. Test object

The municipal website that we evaluated was the site of Deventer (www.deventer.nl), a city in the eastern part of the Netherlands with a population of 97,000. The site is primarily intended for citizens of Deventer, but it also offers elaborate information for those who plan to move there or intend to visit the city. For international visitors there is a short summary, in English or German, of what Deventer has to offer.

As Fig. 1 illustrates, Deventer's home page contains quite a few pictures and bits of information. The information in the middle part of the site is grouped into four categories: companies, visitors, residents and local news. Above these categories there is a welcoming message ('Welcome to Deventer') and at the right-hand, there is a search function. The site's main menu is presented at the left top. This menu, designed to quickly guide the user through the site, has seven options:

1. Take me to…
2. Tell me more about…
3. I'd like to file a complaint about …
4. I live in …
5. Give me the latest news on…
6. Municipal guide
7. Digital service desk

By clicking on each of these general items, a new menu is presented with further, more specific options. For instance, item 1 ('Take me to…') guides users to information on taxes, jobs and income, construction and living environment, the digital service desk, and the registration service. Item 2 ('Tell me more about') has a submenu with links such as 'Facts and figures,' 'Vacancies,' 'Contact details,' 'Tourist information,' and so on. Usually, the submenus of each item lead to one or two additional submenus. This type of information architecture, which is described by Farkas and Farkas (2000) as a "deep" hierarchy, requires substantial navigation. Visitors to the Deventer website first have to navigate through various menus before they are confronted with any kind of text. Reading constitutes a smaller task component than navigating/selecting links in the Deventer site, and takes place only after a substantial bit of navigation.

In this respect, the Deventer site clearly differs from the previous municipal website that we evaluated (Van den Haak et al., 2007). As we indicated in the introduction, this website did include a very substantial reading component. More specifically, it started with a main menu, from which the user had to select a particular link, and then offered informational pieces of text on every subsequent page that helped the user to decide which link to click on next. In other words, the focus here was on reading *before* navigating rather than on reading *after* navigating.

2.2. Participants

Our study involved eighty participants, all students at the University of Twente (Enschede, The Netherlands). These students were gathered by means of printed and e-mail announcements and received a small financial compensation for their participation. About half of them (43 students) were enrolled in Communication Studies. The remaining 37 took different courses. The average age of the students was 21, and there were somewhat more female than male participants (47 vs. 33). At the time of their participation, the students had averaged three years at the university. Nearly all of them (65 students) had some previous experience with a municipal website. However, none of the participants knew the Deventer website. This made the participants a suitable target group as they had experience with the kind of test object that was evaluated, but not with this specific test object. All participants were evenly assigned to the three conditions in the study. The only difference with respect to the demographic details of the participants in the three conditions was that the CTA participants were significantly older (average age of 23) than the participants in the other two conditions (average ages of 21 in the CI condition and 20 in the RTA condition). We checked the correlation between age and the dependent variables in our study, resulting in only a weak correlation (0.29) with one of the problem types (see Processing the data).

2.3. Tasks

To evaluate the Deventer website with the three usability test approaches, we formulated five main tasks, divided into eleven smaller subtasks (See Fig. 2). Each of the main tasks was preceded by a small scenario description, which explained the context and provided relevant details for the task performance (marital status of the subject in the scenario, etc). The tasks were so-called known-items tasks, which meant that they required participants to

1. Registering as a new inhabitant
You've recently moved to Deventer and need to register as a new inhabitant.

1A. *Can you submit a written registration to the city council? Yes/No**
1B. *If yes, what procedure should you follow?*

1C. *Can you submit an online registration to the city council? Yes/No**
1D. *If yes, what procedure should you follow?*

2. Renovating your new house
Your new house needs renovating. As it is placed on the protected property list, you need to carry out the renovations according to various city council rules and regulations.

2A. *Will the city council subsidize the renovations to your house? Yes/No**

3. Relocating a tree
Apart from renovating your house, you also wish to relocate a tree in your garden.

3A. *Will you need official permission from the city council for the relocation? Yes/No**
3B. *If so, are there costs involved?*

4. Paying taxes
Having moved to Deventer (as of 1 May 2005), you would like to know which taxes you have to pay in 2005.

4A. *Do you have to pay user taxes? Yes/No**
4B. *Do you have to pay home ownership taxes? Yes/No**

5. Parking in Deventer
While you don't have a car yourself, your partner, who is not residing in Deventer, 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? Yes/No**
5B. *Are there any (other) parking options that are free?*

indicate which of the two is correct

Fig. 2. Scenario-based tasks designed to evaluate the Deventer website (translated from Dutch).

find information that was known to exist on the municipal website. As such, these tasks could easily be evaluated (both by the participants and the test facilitator) in terms of correctness (Kim, 2001). All tasks could be carried out independently from one another in order to prevent participants from getting stuck after one or two tasks.

Rather than attempting to evaluate the entire Deventer site, 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 Deventer. In this way, we were able to evaluate a very specific and manageable portion of the site and to match the tasks to the participants' real-life situation. After all, as 'outsiders' (inhabitants from Enschede, not Deventer), the students would have fewer difficulties imagining that they would *move to* Deventer than that they would actually *be* a Deventer resident. The ecological validity of the tasks was, as such, ensured.

2.4. Questionnaires

In addition to performing the above tasks, the participants also completed two questionnaires. The first questionnaire, which the participants received on entering the usability lab, contained questions about demographic details including age, gender, and education. It also addressed the participants' previous experience in working with municipal websites.

The second questionnaire measured how the participants had felt about participating in the study. This questionnaire included four aspects: (1) the participants' 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), (3) the participants' evaluation of the tasks that they performed (e.g. 'How satisfied are you with the tasks you performed?', '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 rated their experiences on five-point scales based on semantic differentials. The questionnaire also included some blank space for additional comments.

Participants in the concurrent think-aloud condition (CTA) and the constructive interaction condition (CI) were requested to complete the second questionnaire at the very end of the study, i.e. after they had completed their task performance. The participants in the retrospective think-aloud condition (RTA) were given their second questionnaire in two installments on two occasions. The first installment, containing questions on their method of working, was handed out to them upon completion of their task performance. The second installment, with questions on how the participants had experienced thinking aloud retrospectively, was given to them once the retrospective session had finished.

2.5. Experimental procedure

Our study included sixty sessions, which were all held individually in the same usability lab. Twenty sessions were devoted to twenty CTA participants; another twenty sessions were devoted to twenty RTA participants; and the remaining twenty sessions involved forty CI participants, who participated in the study in teams of two. During each individual session, the movements on the computer screen and

the participants' voices were recorded on video. There was also a facilitator who observed the participants and took notes.

The experimental procedures of the three conditions were exactly the same as the procedures in our previous experiment [reference deleted for review purposes], this to ensure a valid comparison between the present and previous study. For the sake of completeness, a description of these procedures is offered below.

In the CTA condition, the experimental procedure was the following. Upon arriving, the participant filled in the first questionnaire on personal details and previous experience in working with municipal websites. 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 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 municipal website, and not you, who is being tested'. Once the participant had finished his/her task performing, s/he was given the second questionnaire to measure how s/he had experienced her/his participation.

The experimental procedure in the CI condition was as follows. Similar to the CTA condition, the two participants in the CI condition began by filling in the first questionnaire. After completing these questionnaires, the participants were seated randomly at the computer, with one of them sitting in front of it and the other next to it. They then received instructions which explicitly told them to *work together*, as in '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 were told not to turn to the facilitator for assistance. Once their task performance was completed, the participants both received the second questionnaire to indicate how they felt about their participation.

In the RTA condition, the experimental procedure began, once again, with the questionnaire on personal details and prior knowledge. As in the other two conditions, the participants then received the tasks and oral instructions, but now they were instructed to carry out the tasks in silence, again without assistance from the facilitator. Having done that, they had to fill in the first part of the second questionnaire, with questions on their method of working. The participants were then shown a recording of their performance on video and were asked to comment on the process retrospectively. Finally, they received the second part of the questionnaire, designed to measure how they had experienced thinking aloud retrospectively.

2.6. Processing the data

When the sixty sessions were completed we produced transcripts of all the CI, CTA, and RTA verbalizations, and wrote down all the participants' navigations through the municipal website. We then examined these navigations in order to detect usability problems that revealed themselves while the participants were working with the Deventer website. Our criterion for marking a particular situation as problematic was that it should deviate from the optimum working procedure for each task. In addition, we closely studied the transcripts, identifying verbal indicators of problems experienced such as expressions of doubt, task difficulty, incomprehensibility, or annoyance related to the use of the website.

Our analysis of the data collected involved a number of steps. First, we calculated the total number of usability problems detected in each usability method. Next we labeled all problems on the basis of how they had surfaced in the data (i.e. through observation of the behavioral data, through verbalization by the participant, or through a combination of observation and verbalization). Then, two independent coders categorized all detected problems into 9 specific problem types. These types, illustrated in Fig. 3, are based on the categorization that we used in our previous study (Van den Haak et al., 2007). The inter-coder reliability was computed using Cohen's kappa. The overall kappa was 0.91, which indicates a highly satisfactory level of inter-coder agreement.

Apart from the above-mentioned problem types, the participants in our study also experienced occasional technology problems including trouble with the network connection, the browser, or the computer. Some of the problems were not related to the site but to the participants' failure to read the task(s) properly. Neither these technology problems nor the problems that were unrelated to the site were included in our analyses.

Once all problems were categorized into the various problem types we had three independent experts rate each individual problem in terms of relevance. There were two ratings on a 5-point Likert scale. The experts first judged the likelihood of the problem and then, on a separate occasion, its impact (thereby assuming that the problem would indeed be likely to occur) on the proper working of the website. The scores for the likelihood of the problems were then multiplied by the scores for the problems' impact, resulting in a score for relevance. With scores ranging from 1 to 25, we took their square roots as final scores for the relevance of problems.

Next, we used two indicators to evaluate task performance in the three usability methods. We calculated the number of subtasks completed successfully as well as the time required to complete these

1. **Comprehension**:	the participant finds that the information on the site is not clear or applicable; he experiences syntax problems; he finds the choice of vocabulary problematic
2. **Relevance**:	the participant feels that certain information should not be included or should be cut down
3. **Completeness**:	the participant feels that information is missing or more elaboration is needed
4. **Structure**:	the participant finds that the order of information is problematic or that the structure is not clearly signalled
5. **Formulation**:	the participant does not appreciate a particular formulation
6. **Graphic design**:	the participant does not appreciate layout or illustrations
7. **Correctness**:	the participant detects a violation of syntax, spelling or punctuation rules
8. **Data entry**:	the participant does not know how to enter data on the site
9. **Visibility**:	the participant fails to spot a particular link, button, or piece of information on the site

Fig. 3. Classification of problem types.

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

	CTA		CI		RTA	
	Mean	SD	Mean	SD	Mean	SD
Observed	6.3	3.7	9.1	6.4	7.2	4.3
Verbalized	1.9	1.8	2.2	2.2	2.5	1.9
Observed and verbalized	4.9	3.4	5.5	3.2	6.6	4.6
Total	13.1	3.4	16.8	7.0	16.3	4.7

tasks. Finally, to investigate how the participants felt about the tasks they performed, we analyzed their answers to the questions on task performance we posed in the post-session questionnaire. These post-session questionnaires were also analyzed with a view to investigating how the participants had generally felt about participating in this study.

3. Results

This section will discuss the following results: (1) the feedback (number and types of problems) that was collected with the three usability methods, (2) the relevance of the problems detected, (3) task performance (measured in terms of completion time, number of successfully performed tasks, and the participants' own estimation of how successful their task performance had been), and, (4) the experiences of the participants regarding participation in this study.

3.1. Number and types of problems detected

Analysis of the sixty recordings resulted in a total number of 181 different problems. We will first discuss this output by comparing the mean number of problems and problem types detected per session in each condition. Then we will consider the number of different problems detected in each condition and the overlap that exists between them.

Table 1 shows the mean number of problems detected per session. It classifies all problems according to how they surfaced: (1) by observation, (2) by verbalization, or (3) by a combination of both. Anova testing revealed that there were no significant differences between the three usability methods, neither in terms of the total number of problems detected nor in terms of the ways in which these were detected.

To investigate the *types* of problems that were detected in the three usability methods, all problems were labeled according to the problem categorization that we described above. Fig. 4 offers a selection of problems as they occurred in the three conditions.

Table 2 shows the overall distribution of problem types in CTA, CI, and RTA. Anova testing indicated that there was only one significant difference: the RTA participants detected more graphic design problems than the participants in the other two conditions. This difference, however, concerns a very small number of problems (0.2 for RTA against zero problems for the other two conditions), which makes it fair to say that the three usability methods revealed similar problem types in the municipal website. As Table 2 illustrates, the vast majority of the problems detected by the three conditions concerned problems of comprehension. This result can be explained by the fact that in working with the Deventer site, the participants had to find their way through many (sub)menus of links, each of which they had to interpret (often without context) before being able to move on to the next level.

Let us now look at the problems that were detected under each condition (i.e. the list of individual problems regardless of how many times they were detected) as well as the overlap between them. In the CTA condition, 100 different usability problems were detected, while the CI and the RTA condition each revealed 112 different problems. This means that the CTA method was the least fruitful in terms of detecting individual problems.

With respect to overlap in the three lists of usability problems, there were only 46 problems (25%) that occurred in each of the three conditions. The overlap between two rather than three conditions was somewhat higher, ranging from 34 to 36%. As these relatively low percentages indicate, there were a substantial number of unique problems in each condition. The CTA condition revealed 20 unique problems, while the RTA and CI conditions produced 30 and 34 unique problems respectively. These results can be explained by the sheer volume (i.e. the quality and quantity of pages) of the municipal website. Still, if we take the frequency of the problems into account, the degree of overlap was considerable: problems detected in one condition by at least five participants were in 89 to 96% of the cases also detected by at least one participant in one of the other conditions. This means that each of the three methods could clearly detect the main output of the other two.

3.2. Relevance of the problems detected

As was mentioned above, three experts evaluated all 181 individual problems in terms of likelihood and in terms of impact. Rating took place 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 a reasonably reliable scale (Cronbach's alpha = 0.62). With an average score of 2.56, the problems in the current study were rated as moderately relevant.

Regarding the relevance of the problems detected in the three conditions, an analysis involving 95% confidence intervals showed that 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.

With respect to the manner in which the problems were detected, we found that in the CTA condition, the problems that were detected purely by means of verbalization proved more relevant than the problems that were detected by means of observation or by means of a combination of verbalization and observation. The RTA and CI methods showed a similar trend, with verbalized problems being more relevant than problems detected otherwise, but here no significant differences occurred. These (significant) findings cause a paradox: on the one hand, the verbalized problems in this study are clearly judged as valuable; on the other hand, they regrettably form just a small part (15%) of the entire output of each of the three usability methods.

Finally, we also considered the degree of correlation between relevance and frequency of the problems detected. This correlation did not differ between the three conditions, and basically proved non-existent ($r = 0.04$, n.s.).

Table 2
Types of problems detected per session in the CTA, CI and RTA condition

	CTA		CI		RTA	
	Mean	SD	Mean	SD	Mean	SD
Comprehension	11	3.5	14.3	6.6	14.4	4.6
Relevance	0	0	0.1	0.2	0.1	0.3
Completeness	0.5	1	1.2	1.8	0.3	0.8
Structure	0.2	0.4	0.2	0.4	0.2	0.4
Formulation	0	0	0	0	0	0
Graphic design	0*	0	0*	0	0.2*	0.6
Correctness	0	0	0.1	0.2	0.1	0.2
Data entry	0	0	0	0	0	0
Visibility	1.4	1.1	1	0.9	1	0.8

* $p < 0.05$.

1. **Comprehension**:	the participant does not know the meaning of the link 'area 1'
2. **Relevance**:	there is too much information on the web pages
3. **Completeness**:	the Web site lacks a clear overview of parking possibilities
4. **Structure**:	the information on the registration service page is not clearly structured
5. **Graphic design**:	the participant feels that the font used in the digital service desk is too small
6. **Correctness**:	the Euro symbol is presented incorrectly on the Web site
7. **Visibility**:	the participant fails to see the sitemap on the Web site

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

3.3. Task performance

To measure task performance in the three usability approaches we looked at the number of subtasks that were completed successfully as well as the time it took to complete the entire set of tasks. Results are shown in Table 3.

Anova analyses indicated that there were no differences between the three test approaches, neither in terms of successfully completed subtasks nor in terms of the time it took the participants to complete these tasks. Apparently, task performance was not affected by the double workload of having to perform tasks and think aloud simultaneously (in the CTA condition) nor by the instruction to work together (in the CI condition).

As we mentioned earlier in this article, we looked at the participants' estimation of how they performed their tasks by including the following three questions in the post-test questionnaire on participant experiences: (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? The answers to the first two questions had to be indicated on a five-point scale.

As the participants in the CI condition were working in pairs, each with a different role (actor/observer) that may have influenced their experiences, they will be treated as separate subgroups in the analyses of the post-test questionnaire. The actors (the participants working behind the computer) will be referred to as CI Actor, while the observers (those sitting next to the person working behind the computer) will be named CI Co-actor.

The results of the questions on task performance are presented in Table 4. There was only one significant difference between the conditions: the Actors and Co-actors were more satisfied with their performance than the RTA participants (Anova, $F(3,76)=6.21$, $p<0.01$, Bonferroni post hoc analysis, $p<0.05$). Apparently, working in a team has a positive effect on the participants' judgment of the tasks they performed.

3.4. Participant experiences

This section discusses the remaining aspects of the post-test questionnaire:

Experiences with (1) having to think aloud (concurrently or retrospectively) or working together, (2) Method of working, and (3) Presence of the facilitator and the recording equipment. As in the previous section, which discussed participants' estimations of their task performance, the CI Actors and Co-Actors will be considered separately.

The participants were asked how they had felt about having to think aloud (concurrently or retrospectively) or working together by indicating on a 5-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=0.77). Anova testing and Bonferroni post hoc analyses showed that both the Actors (mean score=2.0) and the Co-Actors (mean score=2.1) were significantly more positive about the CI method than the CTA participants (mean score=2.8) and RTA participants (mean score=3.3) about their respective methods (Anova, $F(3,76)=31$, $p<0.01$, Bonferroni post hoc analysis, $p<0.05$). These results are similar to the results of our previous study, which strongly suggests that constructive interaction is evaluated more positively by its participants than the other two usability methods.

Results for the second aspect of the questionnaire (i.e. method of working) were collected by asking the participants to estimate, on a 5-point scale in what respects their working procedure differed from usual (faster or slower, more focused or less focused, etc.). As is clear from Table 5, the scores for all items in all three conditions were rather neutral, indicating that the participants felt that they had not worked very differently from usual. Three of the eight items showed significant differences between the conditions. The CI Actors felt they had been less persevering and had worked less stressfully than was indicated by the self-reports generated by the RTA and CTA participants. The CI Actors and co-actors felt that they had worked faster than was indicated by the self-reports generated by the RTA participants.

The third aspect of the questionnaire involved questions about the presence of the facilitator and the use of recording equipment. Participants were asked to indicate, on a 5-point scale to what extent they found it unpleasant, unnatural, and disturbing to have the facilitator present during the study. They were then asked the same

Table 3
Task performance in the CTA, CI and RTA condition

	CTA		CI		RTA	
	Mean	SD	Mean	SD	Mean	SD
Number of tasks completed successfully	7.4	2.2	8.1	1.4	7.7	1.8
Overall task completion time in minutes	25.5	6.7	24.8	7.0	24.6	5.3

Table 4
Participants' evaluation of the five tasks that they performed

	CTA		CI Actor		CI Co-actor		RTA	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
How satisfied/unsatisfied are you with the tasks performed?	2.8	0.9	2.1*	0.9	2.1*	0.8	3.0*	1.0
How easy/difficult did you think the tasks were?	3.2	0.8	2.8	1.0	3.0	1.1	3.4	0.9
How many tasks do you think you performed correctly?	3.3	0.9	3.4	0.8	3.6	0.8	3	0.9

Note. Scores for the items 'satisfaction' and 'ease of task performance' are indicated on a five-point scale (1 = very satisfied, 5 = very unsatisfied, and so on).
* $p<0.05$.

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

	CTA		CI Actor		CI Co-actor		RTA	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Faster–slower	3.4	0.8	2.8*	1.1	2.7*	1.2	3.6*	0.6
More–less focused	3.0	0.9	2.9	0.9	2.9	0.4	2.4	1.0
More–less concentrated	2.7	0.8	2.8	0.8	2.5	0.8	2.6	0.9
More–less persevering	2.3*	1.0	3.3*	0.7	3.0	0.9	2.4*	1.1
More–less successful	3.1	0.6	3.1	0.6	2.8	0.6	3.2	0.5
More–less pleasant	3.4	0.5	3.0	0.5	3.1	0.9	3.1	0.7
More–less eye for mistakes	2.3	0.9	2.8	1.0	2.8	0.6	2.5	0.7
Less–more stressful	3.4*	0.5	2.7*	0.9	2.9	0.7	3.3*	0.6

Note. Scores for these items are indicated on a five-point scale.
* $p<0.05$.

question with respect to the presence of the recording equipment. Anova testing revealed that there were no significant differences between the three usability approaches regarding either of the questions. Moreover, with average scores ranging between 2.4 and 2.7, the participants clearly felt only marginally affected by the experimental setting.

In summary, the three usability test approaches largely revealed similar results with respect to the participants' experiences in the present study. The few differences that were found indicated that the CI method is most positively evaluated by its participants. This suggests that participants prefer evaluating in teams rather than as individuals.

4. Discussion

As we pointed out in the introduction to this article, the present study was not so much conducted with a view to establishing the importance of usability testing or determining the value of the think-aloud methods within an E-Government context. The importance of usability testing has, after all, long been recognized, and so has the value of think-aloud protocols, a value which has only been further emphasized by the large number of individual problems detected.

What we *were* interested in was to learn whether the three usability methods that we employed for the evaluation of an earlier municipal website would reveal the same (or different) results when applied to a municipal website with a distinctly different information architecture. In the remainder of this section, we will first summarize the results of our present evaluation and then point out, and where possible explain, any relevant similarities and differences between the present and our previous study. In order to facilitate comparison, Fig. 5 offers an overview of the main results of both the present (Deventer) municipal website evaluation and the previous municipal website evaluation.

With respect to the overall usefulness of the three usability methods, the present study revealed a few differences between the methods. The CTA method proved less fruitful in detecting individual problems than the RTA and CI methods, while the latter method received a more positive participant evaluation than the CTA and RTA methods. On the whole, however, each of the methods was equally fruitful in generating output: there were no differences with respect to the average number, type and relevance of problems or the way in which these problems came about. The three methods were also comparable in terms of task performance, revealing no differences regarding task completion times and number of tasks completed successfully.

In revealing largely similar results among each of the three evaluation methods, the present municipal website evaluation is comparable to our previous municipal website evaluation (See Fig. 5). There, too, we found no differences between the three think-aloud

Results in terms of	**Present municipal web site evaluation**	**Previous municipal web site evaluation**
Average number of problems detected	No difference between the three usability test approaches	No difference between the three usability test approaches
Number of individual problems detected per method	CTA proved less fruitful than RTA and CI	CTA proved less fruitful than RTA and CI
Average number of problems according to detection means*	No difference between the three usability approaches	***RTA experienced more observable problems than CI***
Contribution of verbalizations to the output of CTA in %**	52%	***77%***
Types of problems	No difference between the three usability test approaches	No difference between the three usability test approaches
Relevance of problems	No difference between the three usability test approaches	No difference between the three usability test approaches
Successful task completion	No difference between the three usability approaches	***RTA was less successful than CI in completing the tasks***
Range of success rates	67 -74%	***81-94%***
Task duration	No difference between the three usability approaches	***CI took less time than CTA to perform the tasks***
Participant estimation of task performance	CI participants were more satisfied with performance than CTA and RTA	CI participants were more satisfied with performance than CTA and RTA
Participant experiences regarding the usability methods involved	CI was evaluated more positively than CTA and RTA	CI was evaluated more positively than CTA and RTA

Fig. 5. Overview of the main results of the present and previous municipal website evaluations (NB: differences between the two evaluations are marked in bold italics). *Detection means = problems detected by either observation, verbalization or a combination of both. **Verbalizations = problems detected either by means of verbalization or by means of a combination of verbalization and observation (See also the section, Number and types of problems detected). ***Scores indicated on a five-point scale (1 = irrelevant, 5 = highly relevant).

variants in terms of average number, type, and relevance of the problems detected. Taken together, the two evaluations thus seem to suggest that the CTA protocols, RTA protocols, and constructive interaction are in principle interchangeable for the evaluating of municipal websites. This would imply that usability testers who intend to evaluate a municipal website can simply choose any of the three methods at random.

Nevertheless, it would be too early to draw such a conclusion, as we did find a number of differences between the present and previous municipal website evaluation—differences that point to different workings of the usability methods depending on the kind of information architecture (substantial navigating versus substantial reading) of the tested website. As is clear from Fig. 5, the previous evaluation revealed a significant difference with respect to the means of problem detection. The RTA participants experienced more observable problems than the CI participants. Our explanation for this difference was that the RTA participants, in working silently, presumably skimmed rather than carefully read the information that was presented on the middle column of each page of the previous municipal website. Since this information was essential for selecting appropriate links from the right and left columns (see also the section 'Test object'), the RTA participants experienced relatively more observable problems. The fact that this difference concerning observable problems did not occur in the present study can be explained by the fact that the Deventer site has a different information architecture from the previous municipal website. Rather than having to read before being able to properly select links, participants using the Deventer site have to navigate before arriving at any reading material. As such, the RTA participants were less dependent on carefully reading in order to make the right navigational decisions.

A second difference between the previous and present municipal website evaluation concerns the contribution of verbalizations to the output of the CTA method, which was considerably lower in the present evaluation (52%) than in the previous evaluation (77%). This difference in output might also be explained by the different information architecture of both municipal websites. As we suggested in the introduction, navigating, which took a prominent place in the Deventer site, may have involved a heavier workload than reading, which formed the main activity in the municipal website of the previous evaluation. This presumably heavier workload, which is supported by the fact that the Deventer participants were less successful in completing their tasks (success rates ranging from 67–74%) than the participants in the previous evaluation (success rates ranging from 81 to 94%), may have caused reactivity in the CTA method (i.e. may have caused the CTA participants to offer relatively fewer purely verbalized problems and to experience relatively more observable problems).

Two final differences between the previous and present evaluation relate to task performance. The first concerned the CI participants. In the previous evaluation, they took less time to complete their tasks than the participants in the CTA condition. We explained this finding by arguing that by working in teams of two, the CI participants had two pairs of eyes to perform the large reading component that the previous municipal website entailed, which allowed them to locate essential information more quickly. The other significant difference concerned the RTA participants in the previous evaluation, who proved less successful in performing their tasks than the participants in the CI condition. Our explanation for this difference was similar to the one offered for the significantly larger number of observable problems in the RTA condition than in the CI condition (see above): in working silently, the RTA participants presumably skimmed rather than carefully read the information that was presented on the middle column of each page of the municipal website. Since this information was essential for selecting appropriate links from the right and left columns (see also the section 'Test object'), the RTA participants were relatively less successful in completing their tasks.

The fact that neither of these differences relating to task performance occurred in the present study can again be explained in terms of information architecture: on the Deventer municipal website, the CI participants had to navigate substantially before they could engage in reading, which meant that they had lost the advantage of finding information more quickly, and instead spent time on trying out the various navigational paths that they could both propose to follow. Likewise, as the RTA participants had to navigate substantially before they could engage in reading, they were less dependent on carefully reading texts in order to perform their tasks correctly.

From the above four differences, some very preliminary conclusions can be drawn regarding the suitability of the three usability methods for the two types of test object. The participants in the RTA method seemed to perform worse (experiencing significantly more observable problems and performing their tasks less successfully than the CI participants) when faced with the substantial reading task of the previous municipal website than when dealing with the substantial navigating task of the present Deventer website. The CI participants, on the other hand, seemed to work faster when engaging in reading than in navigating. There was some evidence (in the shape of a smaller contribution of verbalized problems to the CTA output) that the CTA participants experienced more reactivity in the Deventer evaluation, while having to think aloud and navigate at the same time, than in the previous municipal website evaluation, where they had to think aloud and read simultaneously. These results offer no definitive support for using the CTA and CI methods for evaluating largely textual municipal websites and the RTA method for largely navigational municipal websites, yet they do indicate that further research into the effect of task type on the workings of the evaluation methods is desirable.

A second aspect to be taken into account in future research includes the possible effect of task formulation on the workings of the CTA, RTA, and CI methods. As indicated in Method section, the tasks performed by the participants in the Deventer evaluation (as well as those performed by the participants in the previous evaluation) were formulated as known-items tasks (tasks requiring participants to find information that was known to be present on the municipal websites). These tasks allowed the participants to evaluate the correctness of their performance, and this may have caused them to work differently than had they been performing open-ended tasks. It is well imaginable, for instance, that the faster task completion time that was found with the CI participants in the previous municipal website evaluation would disappear once they could no longer check whether the information they had found was, in fact, the correct information for completing their task.

Another aspect that might affect the workings of the three evaluation methods involves the characteristics of the participants in a usability evaluation. While we did ensure in both of our evaluations that the people who took part in the studies were evenly divided over the three methods with respect to sex, age, education, and experience with municipal websites, we only considered one specific target group: students. As municipal websites cater to a much broader audience, replication of our studies involving different types of participants would be useful.

A final point for consideration concerns the application of our results to municipal websites in countries other than the one from which our test objects originate. It seems clear that more research on an international (rather than a national) scale is desirable. Our present study hopefully offers a valuable basis for this kind of research.

References

Alavi, S. M. (2005). On the adequacy of verbal protocols in examining an underlying construct of a test. *Studies in Educational Evaluation*, *31*, 1–26.

Chadwick, A., & May, C. (2003). Interaction between states and citizens in the age of the Internet: "E-Government" in the United States, Britain and the European Union.

Governance: An International Journal of Policy, Administration and Institutions, *16*, 271–300.
Choudrie, J., Ghinea, G., & Weerakkody, V. (2004). Evaluating global E-Government sites: A view using web diagnostic tools. *Electronic Journal of e-Government*, *2*, 105–114.
Cullen, R., & Houghton, C. (2000). Democracy online: An assessment of New Zealand government websites. *Government Information Quarterly*, *17*, 243–267.
De Jong, M., & Lentz, L. (2006a). Municipalities on the Web: User-friendliness of government information on the internet. *Lecture Notes in Computer Science*, *4084*, 174–185.
De Jong, M., & Lentz, L. R. (2006b). Scenario evaluation of municipal Websites: Development and use of an expert-focused evaluation tool. *Government Information Quarterly*, *23*, 191–206.
Dumas, J. S., & Redish, J. C. (1999). *A practical guide to usability testing*, Revised edition Exeter: Intellect.
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.
Eschenfelder, K. R. (2004). Behind the Website: An inside look at the production of Web-based textual government information. *Government Information Quarterly*, *21*, 337–358.
Eschenfelder, K. R., & Miller, C. A. (2007). Examining the role of Website information in facilitating different citizen-government relationships: A case study of state Chronic Wasting Disease Websites. *Government Information Quarterly*, *24*, 64–88.
Farkas, D. K., & Farkas, J. B. (2000). Guidelines for designing Web navigation. *Technical Communication*, *47*, 341–358.
Funkesson, K. J., Anbäcken, E., & Ek, A. (2007). Nurses' reasoning process during care planning taking pressure ulcer prevention as an example: A think-aloud study. *International Journal of Nursing Studies*, *44*, 1109–1119.
Griffin, D., & Halpin, E. (2005). An exploratory evaluation of UK local e-Government from an accountability perspective. *The Electronic Journal of e-Government*, *3*, 13–28.
Heeks, R., & Bailur, S. (2007). Analyzing e-government research: Perspectives, philosophies, theories, methods and practice. *Government Information Quarterly*, *24*, 243–265.
Jaeger, P. T. (2006). Assessing Section 508 compliance on federal e-government Websites: A multi-method, user-centred evaluation of accessibility for persons with disabilities. *Government Information Quarterly*, *23*, 169–190.
Kaaya, J. (2004). Implementing e-Government services in East Africa: Assessing status through content analysis of government websites. *The Electronic Journal of e-Government*, *2*, 39–54.
Kim, K-S. (2001). Information seeking on the Web: Effects of user and task variables. *Library & Information Science Research*, *23*, 233–255.
Marcella, R., Baxter, G., & Moore, N. (2003). The effectiveness of parliamentary information services in the United Kingdom. *Government Information Quarterly*, *20*, 29–46.
Nielsen, J. (1993). *Usability Engineering.* Boston, MA: Academic Press.
Pieterson, W., Ebbers, W., & Van Dijk, J. (2007). Personalization in the public sector: An inventory of organizational and user obstacles towards personalization of electronic services in the public sector. *Government Information Quarterly*, *24*, 148–164.
Potter, A. (2003). Accessibility of Alabama government Websites. *Journal of Government Information*, *29*, 303–317.
Rubin, J. (1994). *Handbook of usability testing: How to plan, design, and conduct effective tests.* New York: Wiley.
Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition*, *17*, 759–769.
Schedler, K., & Summermatter, L. (2007). Customer orientation in electronic government: Motives and effects. *Government Information Quarterly*, *24*, 291–311.
Schellings, G., Aarnoutse, C., & Van Leeuwe, J. (2006). Third-grader's think-aloud protocols: Types of reading activities in reading an expository text. *Learning and Instruction*, *16*, 549–568.
Schneider, J. F., & Reichl, C. (2006). Exploring ease in thinking aloud. *Psychological reports*, *98*, 85–90.
Shi, Y. (2007). The accessibility of Chinese local government Websites: An exploratory study. *Government Information Quarterly*, *24*, 377–403.
Taylor, K. L., & Dionne, J. P. (2000). Accessing problem solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, *29*, 413–425.
Ummelen, N., & Neutelings, R. (2000). Measuring reading behavior in policy documents: A comparison of two instruments. *IEEE transactions of professional communication*, *43*, 292–301.
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 den Haak, M. J., De Jong, M. D. T., & Schellens, P. J. (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 den Haak, M. J., De Jong, M. D. T., & Schellens, P. J. (2007). Evaluation of an informational web site: three variants of the think-aloud method compared. *Technical Communication*, *54*, 58–71.
Van Someren, W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method — A practical guide to modelling cognitive processes.* London: Academic Press.
Welch, E., & Fulla, S. (2005). Virtual interactivity between government and citizens: The Chicago Police Department's citizen ICAM application demonstration case. *Political communication*, *22*, 215–236.
Wong, A. T. Y. (2005). Writers' mental representations of the intended audience and of the rhetorical purpose for writing and the strategies that they employed when they composed. *System*, *33*, 29–47.

Maaike van den Haak is a 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 completing her PhD, she also works as a lecturer in English and translation at the Vrije Universiteit in Amsterdam.

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 website evaluation and usability testing, and is currently working on an additional research line on applied research methods in organizational and corporate communication.

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.