Combining Concurrent Think-Aloud Protocols and Eye-Tracking Observations: An Analysis of Verbalizations and Silences

Research Article

—SANNE ELLING, LEO LENTZ, AND MENNO DE JONG

***Abstract*—Research problem:** *Concurrent think-aloud (CTA) protocols are one of the dominant approaches of usability testing. However, there is still debate about the validity of the method, partly focusing on the usefulness and exhaustiveness of participants' verbalizations. The rise of eye-tracking technology sheds new light on this discussion, as participants' working processes can now be observed in more detail.* **Research questions:** *(1) What kinds of verbalizations do participants produce, and how do they relate to the information that can be directly observed using eye tracking? (2) What do eye movements reveal about cognitive processes at times when participants stop verbalizing?* **Literature review:** *Our study replicates an earlier study by Cooke (2010), who used a combination of CTA protocols and eye tracking in a small sample with experienced and highly educated participants to investigate the validity of CTA. Cooke's results suggest that the additional value of participants' verbalizations is limited: at least 77% of the verbalizations referred to things that could be easily observed with eye tracking.* **Methodology:** *We conducted a study in which 60 participants with different characteristics performed tasks on informational websites. During their task performance, they verbalized their thoughts, and simultaneously their eye movements were measured. The resulting think-aloud protocols were divided in verbalization units, which were coded into content types. Silences were registered, and eye movements during these silences were analyzed.* **Results and discussion:** *We found a different distribution of verbalization types than Cooke (2010) reported, with far more verbalizations where participants formulated doubts, judgments on the website, or expressions of frustration. In our study, verbalizations provided a substantial contribution in addition to the directly observable user problems. We measured a rather high percentage of silences (27%), during which participants most often were scanning pages for information. During these silences, interesting observations could be made about users' processes and obstacles on the website. The implication of our study is that we now have a better understanding of the types of verbalizations that a CTA evaluation might generate. Further, we know that relevant usability observations can be made during silences. A limitation is that we do not know yet the influence of specific characteristics of the evaluation setting on the types of verbalizations and silences. Future research should focus on the influence of evaluation settings on the outcomes of an evaluation, in particular, the influence of characteristics of the participants who are involved in the study.*

Index Terms*—Concurrent think-aloud (CTA) protocols, eye tracking, usability, website evaluation.*

Think-aloud protocols are often used for the evaluation of website usability. In an evaluation with concurrent think-aloud protocols (CTA), participants are asked to verbalize everything that goes on in their minds during task performance. The idea behind this method is that we gain insight into the cognitive processes and the obstacles participants experience. There is an extensive amount of research available on the collection of think-aloud protocols. The theoretical foundation of this method is formed by often cited, important cognitive psychological studies by Nisbett and Wilson [1], and Ericsson and Simon [2] on the reliability and validity of verbalizations in relation to mental processes. Subsequently, Boren and Ramey [3] discussed how the goals in the think-aloud practice in usability testing differ from the goals in cognitive psychology and proposed Speech Communication theory as an alternative theoretical framework that better suits usability aims. Many methodological studies about think-aloud protocols in usability testing followed, such as the way participants should be instructed and prompted [4]–[9]; the effects of variations of the method, such as concurrent versus retrospective conditions [10]–[17]; the effects of different task types [18]; and, more recently, the combination of think-aloud protocols and eye-tracking observations [19]–[27]. Yet, many issues still need to be studied in more detail in order to fully understand the merits and limitations of this method.

Manuscript received September 09, 2011; revised May 16, 2012; accepted June 07, 2012. Date of publication August 09, 2012; date of current version August 16, 2012. This paper is based on a research project financed by the Dutch Organization for Scientific Research (NWO). It is part of the research program "Evaluation of Municipal Websites."

S. Elling and L. Lentz are with the Utrecht Institute of Linguistics (UIL-OTS), Utrecht University, Utrecht, the Netherlands (email: S.Elling@uu.nl; L.R.Lentz@uu.nl).

M. de Jong is with Department of Technical and Professional Communication, Faculty of Behavioral Sciences, University of Twente, Enschede, the Netherlands (email: m.d.t.dejong@utwente.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

IEEE 10.1109/TPC.2012.2206190

The combination of think-aloud protocols and eye-tracking observations leads to new perspectives on existing questions and to new research questions and possibilities. A recent contribution to the research on think-aloud protocols combined with eye tracking was made by Cooke [28]. She carried out a CTA evaluation where ten highly educated and experienced participants conducted four search tasks on a website. During task performance, they verbalized their thoughts, while their eye movements were recorded simultaneously. Her research addressed the accuracy of CTA verbalizations, the types of verbalizations participants produced, and users' behavior during silences.

The results of Cooke's study showed that 77% of the verbalizations consisted of reading aloud texts from screen, or reporting actions that the participants performed at a certain moment. This means that a large amount of verbalizations can also be easily observed by analyzing the participants' eye movements. In addition, Cooke's study showed that participants were silent when they experienced cognitive-processing difficulty. At these moments during which important processes took place, the verbalizations did not provide relevant information about the website and the participants' processes and obstacles. We conclude from this study that the small number of possibly useful verbalizations, combined with the silences during relevant cognitive processing moments, raise questions about the usefulness of a CTA evaluation. The benefits of this method seem not what we would hope for, considering the efforts it takes the participants and the evaluator to conduct a think-aloud evaluation. Judging from Cooke's findings, we can conclude that CTA produces almost no results that cannot be found with other methods that are easier to perform for the participant and the evaluator.

However, more research is needed before we can draw firm conclusions about the yield that a CTA evaluation generates in evaluations in various settings. To what extent are the results of Cooke's study illustrative for usability studies that are done in practice? To answer this question, we replicated Cooke's study, in which we aspired to use a more natural research design, with a larger and more representative sample of participants, more than one website as evaluation object, and less eye-tracking equipment-related restrictions during the evaluations. We also examined more comprehensively the silences during think-aloud evaluations. Our research questions were:

RQ1. What kinds of verbalizations do participants produce, and how do they relate to the information that can be directly observed using eye tracking?

RQ2. What do eye movements reveal about cognitive processes at times when participants stop verbalizing?

Hence, in this paper, we replicate the study of Cooke, addressing research questions on the distribution of types of verbalizations and the cognitive processes during silences. This will lead to more knowledge on the merits and limitations of CTA.

In this paper, we first discuss relevant literature, focusing, in particular, on Cooke's earlier study. After that, we explain the methodology of our study, then present the results of our study, and close by presenting the conclusions and limitations of the study, and suggesting future research.

Literature Review

In this section, we discuss the relevant literature related to our study. In the "theoretical orientation" section, we describe the methodological contribution that our study is aimed toward. Then, we discuss theory on the "accuracy of verbalizations compared to eye movements," "types of verbalizations," and "users' behavior during silences."

Theoretical Orientation Our theoretical orientation is on the methodology of website evaluation with CTA. Combining CTA with eye tracking can provide new insights into the processes behind this method and the results it generates. First, the availability of eye tracking extends and refines the observable parts of participants' problem-solving processes, which raises new questions about the nature and added value of participants' verbalizations. This relates to discussions about the usefulness of CTA protocols. Second, eye tracking offers the possibility to analyze what participants are doing when they do not verbalize their thoughts. This relates to discussions about the exhaustiveness of participants' verbalizations.

Selection of Literature We have selected literature on the following themes: CTA protocols, CTA combined with eye tracking, types of verbalizations in CTA, and users' processes during think-aloud evaluations. The studies selected for our literature review focus explicitly on the

methodology of think-aloud protocols, instead of only reporting evaluations of applications using think aloud. As we perform a replication of Cooke's study [28], this paper forms the starting point of our literature review. The next three subsections discuss Cooke's three main findings.

Accuracy of Verbalizations Compared to Eye Movements According to Cooke, the accuracy of CTA verbalizations can be measured by analyzing the correspondence between users' verbalizations and their eye movements during these verbalizations. Cooke's [28] results on accuracy showed that verbalizations corresponded to eye movements 80% of the time, which would indicate that concurrent verbalizations are highly accurate when compared to eye movements. Cooke explains the 20% discrepancy between the two by the observation that participants often verbalized at a slower rate than their visual processing of the information. As a result, they sometimes verbalized thoughts about one part of the screen, while their eyes were already focused on another part.

In our opinion, however, measuring the accuracy of CTA verbalizations is more complicated than it may seem at first sight. If we try to measure accuracy, we actually want to determine to what extent verbalizations correspond to users' thoughts during task performance (connection 1 in Fig. 1). The comparison between verbalizations and eye movements is based on the idea that eye movements provide insight into users' minds, the so-called eye-mind hypothesis [29] which states that there is a direct relationship between what people are looking at and what they are thinking (connection 2 in Fig. 1). Hence, the assumption is that the eye movements represent thoughts. Consequently, the correspondence between eye movements and verbalizations (connection 3 in Fig. 1) should provide information about the extent to which verbalizations are accurate reproductions of users' thoughts.

However, there are three reasons why assessing the accuracy in this way is problematic. First, let us assume that the connections in Fig. 1 indeed work in the way we described and then look at the largest group of verbalizations in Cooke's study: the verbalizations in which users were reading something aloud from the screen. These verbalizations are deemed accurate by definition because reading aloud requires looking at the screen and verbalizing at the same time, which results in a direct relation through connection 3. Also, connection 2 seems correct, as reading

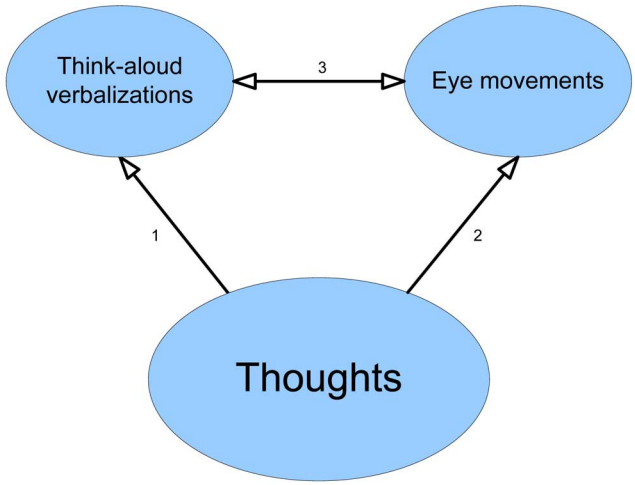

Fig. 1. Relation between thoughts, verbalizations, and eye movements.

aloud requires thinking about the text that is being read. Therefore, the claim seems justified that reading verbalizations are accurate reproductions of thoughts. However, users sometimes might read information aloud without really processing it. In our usability studies, we have seen participants reading aloud a task instruction and then ask permission to read the same text in silence because they did not have a clue of the content they had just read. This means that all reading verbalizations are registered as accurate, while some of them do not necessarily correspond to users' thoughts. Hence, even in situations in which the connections between verbalizations, eye movements, and thoughts seem obvious, we cannot draw firm conclusions on the accuracy of verbalizations.

Second, there are verbalizations that are described as inaccurate because they do not correspond to eye movement observations (connection 3). Research shows that users process information at a faster rate than they can verbalize, such as [30]. Guan et al. [23] explained omissions in retrospective think-aloud data by the different data densities and levels of abstraction of the two methods. Eye tracking provides high density and low abstract level sequence data, while verbal reports have a lower density and a higher abstract level. This means that one verbalization might summarize longer sequences of eye movements and can therefore not be directly related to one specific eye movement, for example in the verbalization "I will now scan this home page" followed by a series of eye movements during which the homepage is scanned. In this example, connection 1 between thoughts and verbalizations is adequate. Also,

connection 2 between the user's thoughts and eye movements seems correct. Yet, Cooke would probably describe the verbalization as inaccurate even though it provides very useful and adequate information about the participants' thoughts.

Third, we think that the assumption of a direct relation between thoughts and eye movements (connection 2) is problematic. Many types of thoughts, such as expectations, comments on missing information, expressions of doubt and confusion, or observations on the user's own behavior in relation to the website, as in "I feel foolish working at this site" cannot be directly related to eye movements. Verbalizations of these thoughts will therefore be described as inaccurate, while they are actually correct manifestations of thoughts. Moreover, it is precisely these types of verbalizations that are often very useful, as they provide relevant information concerning the problems experienced on the website.

To conclude, the relation between verbalizations, eye movements, and thoughts is rather complicated. Both verbalizations and eye movements are limited expressions of the users' thoughts. Therefore, we think a comparison between these methods does not lead to an adequate definition of the verbalizations' accuracy, which is why we will not replicate Cooke's study on accuracy. Combining think-aloud protocols and eye-tracking observations can, however, provide more insight into the types of thoughts and processes that can be visualized by both methods. In our study, we will therefore focus on the types of verbalizations, and the processes that can be observed during silences.

Types of Verbalizations The second research question in Cooke's study addressed the types of verbalizations that the CTA method yields. A content analysis of the verbalizations showed that a majority of 58% of the verbalizations fit into the *reading* category: words, phrases, and sentences that participants read directly from the screen. In 19% of the verbalizations, participants reported *procedures*: their activities on the website, for example what they were looking at. The remaining verbalizations concerned *observations* about the website or participants' own behavior (10%), *explanations* about the rationale or motivation for actions (5%), and *other* verbalizations (8%), which, for example, included incomplete sentences. What can be concluded from this distribution of verbalizations? The *reading* and *procedural* verbalizations, both classified as low-level verbalizations by Ericsson and Simon [2], accounted for 77% of the CTA content, which corresponds to studies of Bowers and Snyder [10], and Zhao and McDonald [9] on types of verbalizations in CTA. Also, Van den Haak, De Jong and Schellens [15] found that participants often verbalized what they were doing physically at a particular moment. Cooke considered these low-level verbalizations as a useful account of a user's on-screen behavior. She argued that *procedural* verbalizations are useful because they provide insight into users' behavior that would not be readily accessible by observation alone. Indeed, some *procedures*, such as descriptions of what users are looking at, are not observable from screen and mouse observations. However, eye-tracking recordings considerably increase the range of possible observations [29]–[32]. Using adequate eye-tracking recordings of a participant's task performance, all *reading* and *procedural* verbalizations can also be observed through participants' eye movements.

This raises the question to what extent the verbalizations add valuable information that cannot be found by looking at eye-tracking observations. Only 15% of the verbalizations in Cooke's study fall into the categories *observations* and *explanations*, which is a very small percentage as only verbalizations in these categories refer to thoughts that might be harder or impossible to observe. If this distribution is widespread, we would question the usefulness of verbalizations in relation to what can be observed. It is doubtful whether all the efforts of carrying out a think-aloud evaluation, and the possible drawbacks of thinking aloud on the ecological validity [15] are worth this 15% yield. However, more research is needed to find out to what extent this contribution of verbalization types is valid.

Cooke's participants were ten highly educated experienced participants who performed rather easy search tasks. They did not verbally report difficulties in completing the tasks, which indicates that the tasks were not cognitively complex for them. However, most usability studies have the purpose of measuring performance of less experienced and less highly educated participants who possibly experience serious problems on websites and, therefore, may verbalize other types of thoughts than Cooke's participants. Van den Haak et al. [12]–[16] showed that more difficulties during task completion may lead to less verbalizations and more observations. We have therefore replicated Cooke's study in a different, more natural, setting,

and we have analyzed to what extent we found the same distribution of verbalization types.

Users' Behavior During Silences The third research question in Cooke's study was an exploratory question concerning users' behavior when they fall silent or use verbal fillers. Cooke found that silences and verbal fillers occurred 16% of the time. These silences may point to cognitive-processing difficulty during task completion. When participants are cognitively occupied with information processing, they seem to prioritize this over their verbalization task. Switching between two tasks is difficult for users, because it requires a reconfiguration of mental resources. Monsell [33] provides an overview of studies on task switching which, for example, point to longer response times and a higher error rate. Users are inclined to concentrate on one task and neglect other tasks, which is called cognitive lockup [34]. This is also visible in evaluations where participants are asked to provide feedback [35]. There are two possible consequences of cognitive lockup during thinking aloud. First, in situations where users need more cognitive energy for task completion, they have problems verbalizing their thoughts, which results in more silences [3]. Second, task performance deteriorates in concurrent think-aloud conditions, because participants use their cognitive energy to think aloud, as was reported for example by Van den Haak et al. [13]. In our study, we will revisit this question on silences in the setting we described before. We will look more precisely at users' actions during silences and thereby further explore what eye movements reveal at times when users stop verbalizing their thoughts. The answer will deepen our understanding of what kinds of information we miss when we only examine the CTA output: the users' verbalizations. We expect that silences will occur more in our study than in Cooke's study, as we will include lower educated participants and more difficult tasks, which probably results in more cognitive-processing difficulty and, therefore, more silences. The exact behavior during these silences has not been explored yet, which means that we have no expectations about the distribution of processes the eye movements will reveal.

METHODOLOGY

In this section, we describe the methodology we have used in our study. As a reminder, the questions underlying this study included:

RQ1. What kinds of verbalizations do participants produce, and how do they relate to the information that can be directly observed using eye tracking?

RQ2. What do eye movements reveal about cognitive processes at times when participants stop verbalizing?

This section starts with an explanation of the choice of research methodology. Then, we describe the participants, and data collection and analysis techniques.

Choice of Research Methodology In our study, the CTA methodology is the central research object. We have used CTA for the evaluation of websites, and we have analyzed the participants' output in order to answer the question on the types of verbalizations the method generates. This distribution might be the same as in Cooke's study, which would enhance the doubts about the usefulness of CTA. However, the more natural setting of our study might also lead to another distribution of verbalizations. We have combined CTA with eye tracking, as eye movements can provide detailed information about the users' actions during silences in the think-aloud protocol. These silences might occur more than in the study of Cooke, as we expect our participants to experience more cognitive load. We have no concrete expectations about the types of actions during silences, because silences during CTA have not been analyzed systematically before.

Participants Sixty participants took part in this study. We obtained the university's approval to conduct human-subjects research, before the study was carried out. The participants, who received financial compensation for taking part in the study, were recruited by a specialized agency that has an extensive database of potential research participants who all signed an approval form to participate in human-subjects research. All participants in our study indicated that they used the internet at least once a week. We aimed at an equal spread of participants over age groups (18–29, 30–39, 40–54, and older than 55), gender, and educational level. We distinguished three education levels, based on the highest form of education that participants had completed. The lowest educational level ranged from elementary school education to junior general secondary education; the medium-level educational group had intermediate vocational education, senior general secondary education, or preuniversity education; the high-level educational group had completed

higher vocational or university education. We aimed at a distribution of characteristics in such a way that, for example, all age groups consisted of equal numbers of men and women and of different educational levels.

Data Collection This section describes, in detail, the data-collection procedures used in this study, information about the tasks to be accomplished, and the procedure used to collect data about the tasks.

Tasks to Be Accomplished: We have replicated Cooke's study on three different websites, on each of which 20 participants conducted three subtasks. Each task had a theme, which was introduced by a scenario which explained the context and provided the relevant details. The tasks differed per website, but they were comparable regarding their difficulty and the length of the optimal navigation path. All tasks covered realistic activities that correspond to what users usually do on municipal websites. The tasks included searching for information, as well as reading, understanding, and applying the relevant information to the described scenario. For one of the websites, for example, the task involved a subsidy that the government offers to support first-time home buyers. Participants needed to search for information on three subquestions: (1) What is the name of this subsidy? (2) Do you meet the requirements for this subsidy? and (3) What should you do to make a request for this subsidy? The shortest navigation path to this information consisted of six links, the last of which was a link to a PDF file with information related to all kinds of possible subsidies for buying or renovating a house. Due to the information in the PDF file, this task on website 1 required somewhat more reading than the tasks on websites 2 and 3. The task on website 2, on the contrary, required little reading but included searching for information about school holiday periods on an inconveniently arranged webpage that contained many rows of holidays per school type and year.

The tasks were performed on one of three websites of medium to large Dutch municipalities: (1) www.apeldoorn.nl; (2) www.dordrecht.nl; or (3) www.nijmegen.nl. The main objective of these types of websites is to provide information and services to their citizens and to other interested users, such as tourists and businesses. These websites invariably contain a variety of information, as they are designed to satisfy the informational needs of a broad target group. The three municipal websites are comparable with regard to the information they cover, but their content is structured and presented differently. A survey evaluation of the three websites resulted in comparable user-quality judgments on navigation, content, and design [36].

Procedure: Data were collected using a remote Tobii 1750 eye tracker with Tobii Studio 1.2 software. A flat monitor was used to display the websites, with a resolution of 1024×768 pixels. The eye tracker used near-infrared diodes to measure eye movements, and was therefore nearly unnoticeable for the participants. Also, the participants had considerable freedom of head movement, which made the test environment rather natural. Video and audio data were captured together with the eye and mouse movements. Our equipment setting was more natural than in Cooke's study, where the eye-tracking method posed some limitations on the study. Participants in her study needed to keep their heads still, which increased the artificial nature of the evaluation.

All 60 evaluation sessions took place individually and lasted approximately 20 minutes each. Participants were told that we wanted to know more about the user friendliness of the municipal websites. We explained that they would perform some tasks on a website, about common things that people might do on these websites. We emphasized that the website was the object of the study and not the participant. In this way, we presented the study as a usability test of the website, and made sure that participants were unaware of our focus on the CTA methodology.

Subsequently, participants were instructed to think aloud during their task performance. They were instructed as follows:

> Please, think aloud during performing the tasks on the website. Keep on verbalizing everything that goes on in your mind. Please, pretend that I am not there and do not ask me for help. Just speak for yourself and verbalize your thoughts. If you are silent for a while, I will remember you to keep thinking aloud. You can say anything that comes into your mind. Remember that it is the website that is being tested and not you.

After this instruction, we emphasized again that participants should verbalize all of the thoughts they had in mind, not just the actions they performed on the website.

After these think-aloud instructions, participants took a seat in front of the computer. We explained that the tasks would appear on the screen and that

the same task descriptions were also available on paper, in case they wanted to re-read something. We asked them to act as if they were working at home and wanted to find the requested information. Next, their eyes were calibrated and we briefly repeated the most important instructions. Then, we started the session by displaying the first task on the screen.

During the user's task completion, the test administrator was in a test room with an open connection to the participant's room. After 5 s of silence, participants were reminded to think aloud with neutral prompts, such as "Please keep thinking aloud." Contrary to Cooke's study, in which prompting was rarely necessary, we regularly prompted participants to keep thinking aloud. If participants had found the information on the website, they could tell the answer to the administrator.

Data Analysis In the analysis stage, all data sources could be replayed simultaneously. Our equipment produced synchronized data with the eye movements, cursor movements, page switches, and audio recordings of participants' verbalizations. This was an improvement compared to the analysis of Cooke, during which the file with on-screen actions and verbalizations needed to be synchronized with the eye-tracking file.

We first analyzed the task performance time per participant, and used analysis of variance (ANOVA) to measure differences on participants' mean task time between the three websites. The next step was the transcription of the verbalizations per participant into protocols. Verbalizations were divided into units which could include single words, but also clauses, sentences, and phrases. Unit borders were determined by pauses between verbalizations and by the content of these verbalizations, following the procedure used by Cooke [28] and Eveland and Dunwoody [37]. We chose to use the term "verbalization units" instead of "thought units," because, in our opinion, verbalizations are manifestations of thoughts and not necessarily thoughts themselves.

In order to answer RQ1 on the types of verbalizations, all verbalization units were coded into seven content categories, of which the first four categories were the same used by Cooke in her study [28].

(1) *Reading*: verbalizations of link labels and parts of texts that participants read directly from screen.
(2) *Procedure*: descriptions of past, current, or future actions (such as "I now click on this link").
(3) *Observations*: verbalized judgments concerning the website or the participants' actions (such as "This web page is not conveniently arranged," or "I don't understand the meaning of this link label.")
(4) *Explanations*: verbalizations in which participants explain or motivate their actions (such as "On this page I expected to find information on buying a house for the first time.")
(5) *Task-related*: verbalizations concerning the tasks the participants were doing, for example on the task description or answers to the scenario questions (such as "Let's see, what was the exact price of the house I want to buy?" or "I think the answer is that you can only register online if you have a scanner".)
(6) *Fillers*: verbalizations when participants did not seem to know what to say, but wanted to break the silence or to meet the requirement to think aloud (such as "okay," "let's see," or "well"). Cooke chose to analyze these fillers as silences, but we consider them to be verbalizations. They can be meaningful, for example, as expressions of doubt or reactions to prompts to think aloud.
(7) *Other*: verbalizations that did not fit into one of the six other categories.

All verbalization units were categorized into one of the seven content types by two independent coders. The inter-rater reliability, measured with Cohen's kappa, was 0.62, which means that there was substantial agreement between the two coders. Subsequently, the first author analyzed all verbalization units on which the two coders did not agree, and chose one of the two content categories in consultation with the two coders. This resulted in the distribution reported in the results. Differences between the websites on the total number of verbalizations were analyzed using analysis of variance (ANOVA). A Pearson's chi-square was used to analyze interaction effects between task performance time and numbers of verbalizations. Differences between websites on the distribution of types of verbalizations were also measured with a Pearson's chi-square test. We did not perform analyses on differences between types

of participants, because the set of 60 participants was too small for such analyses.

To answer RQ2, we analyzed the eye movements during participants' silences. Eye-tracking movements, however, were not recorded adequately for 8 out of the 60 participants, due to (partial) loss of calibration (for two participants on website 1, and three participants on websites 2 and 3). Therefore, these participants were excluded from the analyses for RQ2.

For each participant, all silences longer than 1 s were registered, and the percentage of time that participants were silent was computed. Then, ANOVA was used to determine to what extent the percentage of silences differed between the three websites. Next, we observed for each silence, what the eye movements showed us about the participants' actions during that silence. We distinguished three types of actions: scanning, reading, and fixating. During scanning, fixations move over a page irregularly [25]. Several types of scanning can be distinguished, from globally orienting a new webpage to scanning rows of links in navigation menus. A reading pattern was indicated by short fixations following the text flow horizontally from left to right [25]. Something was considered as reading if at least three fixations and two saccades could be observed in this pattern. Something was classified as a fixation when a participant fixated for a longer time on an element on the page. We have manually analyzed the distribution of these types of visual behavior during the silences.

RESULTS

In this section, the results of the study will be described. The section starts with a description of "who participated in the study," then we will address the "general results" and the results on our two research questions about (1) "types of verbalizations" and (2) "silences in think-aloud protocols."

Who Participated in the Study? A number of 60 participants took part in the study. The participants' ages ranged from 18 to 83 years, with an average age of 40. Table I shows the distribution of participant characteristics concerning age, gender, and education level.

TABLE I
CHARACTERISTICS OF PARTICIPANTS IN THE STUDY

Characteristics	Number of participants (N = 60)
Age:	
18-29	16
30-39	16
40-54	15
> 55	13
Gender:	
male	29
female	31
Education:	
low	18
middle	18
high	24

General Results The mean task performance time per participant differed significantly for the three websites ($F(2,57) = 8.48$, $p < 0.01$). On website 1, the mean task time was 6.37 min; on website 2, this was 3.55; and on website 3, the time was 5.11. A post hoc Scheffé test showed that participants on website 1 had a significantly longer task performance time than participants on website 2 ($p < 0.01$). The participants produced a total number of 2502 verbalization units on the three websites: 712 on website 1, 621 on website 2, and 1169 on website 3. There was a difference between the three websites in the total number of verbalizations ($F(2,57) = 13.10$, $p < 0.05$), with significantly more verbalizations provided on website 3 than on websites 1 ($p < 0.01$) and 2 ($p < 0.01$). In addition, we explored whether a longer task performance time resulted in more verbalizations. This appeared not to be the case ($r = 0.18$, $p = 0.17$).

Types of Verbalizations RQ1 concerned the distribution of types of verbalizations. We distinguished the categories *reading*, *procedure*, *observation*, *explanation*, *task-related*, *filler*, and *other*. Table II shows the overall distribution of the verbalizations over the content categories for the three websites. There are significant differences between the websites on three categories: *reading*, *procedure*, and *other* ($\chi^2(12) = 34.53$, $p < 0.01$). For these categories, the range of percentages is shown in Table II. We will discuss each verbalization category in detail, examine the content of the verbalizations, and compare them to the distribution reported by Cooke [28]. Then, we will discuss the explanations for differences between the two studies, and examine the differences between the three websites in our study in more detail.

In the current study, 23% (581) of the verbalization units can be categorized as *reading*. This percentage differs considerably from the 58% reported by Cooke. The 423 *procedure* verbalizations accounted for 17% of the CTA verbalizations, which is roughly

TABLE II
TYPES OF VERBALIZATIONS

	Current study	Cooke's study
Reading	23% (range 18–27%)	58%
Procedure	17% (range 14–20%)	19%
Observation	34%	10%
Explanation	7%	8%
Task related	14%	-
Filler	3%	-
Other	2% (range 0.6–2.2%)	5%

equal to the 19% in Cooke's study. Participants often verbalized what they were doing on the website, for example "I will click on the 'moving' link." They also reported activities they just had performed, such as "I scanned the homepage for information about buying a house," or activities they planned to perform, such as "I am going to read this text to search for answers on subsidies."

The largest group of verbalizations (850, 34%) concerned *observations* on the website or the participant's own behavior. This substantially differs from Cooke's study, in which this percentage was only 10%. Observational verbalizations in our study sometimes contained neutral descriptions of elements on the website, such as "Up there I see some links to forms" or "What I see here is online procedures on the right side and renting on the left side." Also, participants verbalized expectations, such as "There might be something useful under online procedures" or "The housing link seems logical." Sometimes these expectations turned out to be wrong, illustrated by verbalizations such as "This is not what I looked for" or "I see here that this is not the procedure for online registration." Many *observations* contained a judgment about something on the website, such as "The font is too small here" or "This web page contains a lot of unclear information." Participants frequently expressed their feelings or experiences in relation to the website, for example "I am not sure where to go from here" or "I don't think I can find the information I need." And sometimes they really judged themselves, expressed in verbalizations such as "I clam up completely" or "I feel like a stupid idiot."

The percentage of *explanations* was more or less the same in both studies: 7% in our study (166) and 8% in Cooke's study. These *explanations* were often linked to *observation* or *procedural* verbalizations, for example in "I click on the back link again, because I did not find the information I expected on this page."

A number of 361 verbalizations (14%) were *task related*, such as answers on the scenario questions. Cooke did not include this category in her study. The verbalization of *task-related* comments can be explained by the rather complex and comprehensive scenario tasks we used, which were also presented on paper during the task. This caused situations where participants looked on the paper during task processing and verbalized something about doing this, such as: "Let's see on the paper ... can you register yourself online?" Moreover, we asked participants to answer the questions orally, which also generated verbalizations on tasks, such as "The answer is that I meet the requirements for the subsidy." We considered leaving these *task-related* verbalizations out of the data set, but we have chosen to incorporate them for two reasons. First, ignoring them would suggest that participants verbalized fewer thoughts, while these verbalizations were present and prevented them from verbalizing other things at the same time. Second, the *task-related* verbalizations were mixed up with the other types of verbalizations and occurred during all phases in the process. Even in the answering stage, *task-related* verbalizations alternated with, for example, *reading*, *procedures*, and *observations*.

The 81 *filler* verbalizations accounted for 3% of the total. Participants were frequently stimulated to keep verbalizing their thoughts, which often led to verbalizations meant to break the silence, without really expressing thoughts, such as "let's see," "let's think," "okay," and "eeeeeeh." Sometimes, participants used these *fillers* as a 'bridge' to more elaborate verbalizations that followed a few seconds later. It seems that they needed some time to arrange their thoughts before they could verbalize them.

The *other* category consisted of 40 (2%) verbalizations, versus 5% in Cooke's study. This category contained many uninterpretable and unfinished utterances. Participants also commented on their troubles with thinking aloud, often as a reaction to a prompt to keep thinking aloud.

The comparison of the distribution of the verbalizations over the different categories, shows that the distribution differs substantially between Cooke's study and our study. The most striking differences can be found in the *reading* and *observation* categories: *reading* was, by far, the largest category in Cooke's study, while in the current study, the group of *observations*

was largest. This distribution has consequences for our doubts on the value of the think-aloud verbalizations. In Cooke's study, at least 77% of the verbalizations strongly corresponded to eye movements, which raised the question to what extent do the verbalizations have an added value compared to the observations. In our study, however, only 40% of the verbalizations fall into these more easily observable groups. Moreover, our participants produced many verbalizations in the categories *observation* on website or own behavior and *explanations*, which cannot be directly observed with eye movements. This means that in our study, the participants' verbalizations clearly add to the observations. This result was found on each of the three websites in our study.

How can these differences in the distribution of types of verbalizations be explained? The most important explanation can be found in the improved research design of our study, regarding the participant characteristics, task complexity, and websites. Cooke had ten highly educated and experienced participants who performed search tasks on a website with a rather simple, consistent design. Her participants did not verbally report difficulties in completing the tasks, which indicates that the tasks were not cognitively complex. This is not representative for the evaluation settings in usability practice in which the goal is to find obstacles of real users on real websites. In our study, a great variety of participants performed rather complex tasks on large municipal websites which contained many menus, links, pictures, and pieces of information. Participants in our study often reported difficulties and doubts. All of these observations on complex aspects of the website and on the problems they experienced explain why the *observation* category is so much more extensive in our study than in Cooke's study. This also partly explains why the *reading* category is larger in Cooke's study. However, in our study, participants also verbalized many *reading* units. We had not instructed them to read aloud everything they read, we only asked them to verbalize their thoughts. As a result, participants differed in the extent to which they read aloud, varying from reading everything they saw, to keeping silent for a long time during reading, or just verbalizing some highlights from the text. Cooke's test object contained large amounts of text, a lot of which her participants apparently read aloud, regardless of whether or not they were instructed to do so in the think-aloud instructions.

Although the distribution of the *reading* and *observation* verbalizations differed substantially from Cooke's study on all three websites in our study, we also found some differences between the three websites in our study ($\chi^2(12) = 34.53$, $p < 0.001$). On website 1, the *reading* category was significantly larger than on website 2: 27% versus 18%, while website 2 had a larger amount of *procedure* verbalizations with 20% versus 14% on website 1. This result can be explained by the differences in task-website interaction. One of the tasks on website 1 required reading and applying information from a comprehensive and rather complex text. This might have resulted in more *reading* verbalizations, while on the other website, users performed more actions and verbalized these in the *procedural* verbalizations. Website 3 had significantly more verbalizations in the *other* category, which can be attributed to one participant who produced nine verbalizations during task processing in which she described her trouble thinking aloud, for example "I find it very hard to search for information and to think aloud at the same time. That does not easily go together for me, but I'll try." However, the differences between the three websites are negligible compared to the differences between the results of Cooke's and our study.

The conclusion that can be drawn from the results on RQ1 is that a CTA evaluation can generate verbalizations that actually contribute to our knowledge about the users' processes and obstacles on a website. The concerns that Cooke's study raised about the balance between costs and benefits of CTA, are not confirmed by our study. On all three websites, CTA showed a clear added value with many verbalizations about processes and obstacles that could not be observed with observational methods. The settings of our study reflected, to a large extent, the usability practice, where participants with varying characteristics perform complex tasks on real websites. We therefore expect that also, in practice, the CTA method will show its value for measuring the websites' quality.

Silences in Think-Aloud Protocols The eye movements of 52 participants during silences were analyzed, in order to find out what participants do when they are not verbalizing their thoughts (RQ2). The mean percentage of silences during a session was 27% ($SD = 19\%$) of the total task performance time. There was a high amount of variation between participants, with the percentage of silences observed ranging from 4% to 83%. No differences were found across the three websites in the percentage of time that participants were silent ($F(2,49) = 0.34$, $p = 0.71$). The mean percentage of

TABLE III
ACTIONS DURING SILENCES

Action	Percentage* (range, *SD*)
Scanning	62% (27–93%, *SD* = 15.7)
Reading	24% (0–70%, *SD* = 15.2)
Fixating	15% (0–43%, *SD* = 11.2)

*Note: we report percentages of the total time that participants were silent and not of the total task-processing time.

silences is considerably higher than the 16% that Cooke reports in her study. One reason for this difference can probably be found in participants' experience with thinking aloud. All participants in Cooke's study had been subjects before in at least one usability study that included CTA, whereas none of the participants in our study had any previous experience with thinking aloud. Also, research has shown that thinking aloud is more difficult when the cognitive load on other processes increases, such as [3] and [13], the so-called cognitive lockup [34]. The task complexity in our study, combined with the participant characteristics and the comprehensiveness of the website, might have led to a rather high cognitive load which thus increased the number of silences.

The aforementioned results show that we miss information on approximately a quarter of the time spent on a website when we only rely on participants' verbalizations. What can we observe from participants' actions during these silences? We have distinguished three types of actions: reading, scanning, and fixating, which are shown in Table III.

The first type of actions we observed were scanning processes, which occurred more than half the time that participants were silent (62%). The types of websites in our study contained a lot of information, often with several menus, pictures, and pieces of text on one page. When participants navigated a website searching for information, they were intensively orienting themselves on webpages and scanning these pages for information. These search and scan processes advance very quickly and require much cognitive energy. It is therefore difficult for users to verbalize these processes [34]. This difficulty can further increase if users encounter problems during this process that add to the cognitive load. Probably, it makes no sense to ask participants to verbalize their scanning processes, because it is too difficult (or even impossible) and there is a great risk of disruption of their natural processes. Moreover, we can easily observe the scanning processes when we measure the eye movements. These observations are an important addition to the verbalizations, because silences during scanning processes often contain relevant information about problems that participants experience; problems that would not always be found when we only depend on users' verbalizations. Observations can, for example, reveal users' doubts about what link to choose or what information to use. This is shown by eye movements switching between several objects, for example, links. Often, during these moments, the mouse cursor does not move and users do not verbalize their scanning behavior, which means that these doubts are only observable using eye tracking. This finding has also been reported in other studies, such as [20] and [38].

Another 24% of silences concern reading something on the screen, the second process we distinguished. Reading is a process that can be verbalized by users. In our study as well as in Cooke's [28], participants often read texts aloud. So although reading processes can, in principle, be verbalized, participants sometimes chose to read in silence, for example, because they did not think it necessary to read everything aloud. For some participants in our study, it seemed difficult to read something aloud and to process the content at the same time. For these people, reading aloud required too much cognitive energy. From an evaluation perspective, we think that it is no problem if users choose to read in silence. Reading processes can be easily observed, so we do not miss any information about the users' processes. Problems related to understanding the content can be found by observing users re-reading parts of text, or by verbalizations about the content that are made afterwards, for example, wrong paraphrases of the content or comments on comprehensibility.

The third process, longer fixations on objects, was observed in 15% of the silences. We suppose that participants were processing information during these fixation moments, for example, trying to relate the object to their search goals or because they were in doubt what to do. These fixations can, in principle, be verbalized. However, the cognitive processing that goes on during these fixations often seems to require so much energy that participants are not able to verbalize their thoughts, and fall silent. These silences can sometimes point to problems, and observations of eye movements can be used to find out what users are looking at exactly. However, additional verbalizations are needed to determine if there really is a problem and to define the essence of this problem.

The aforementioned three processes during silences are sometimes verbalized immediately before or after silences, when participants give an account of something they will do or have just done during a silence. Examples are: "I will now search for links that have something to do with buying a house" or "I was reading through this text and thinking on how to register myself as a new inhabitant." Participants can also scan or read something in silence and then verbalize a conclusion about what they just saw. This conclusion can be positive ("I think I have found relevant information on school holidays here"), neutral ("This extract contains a lot of information on subsidies"), or negative ("No, this information is not relevant to me"). In Cooke's study, these verbalizations would have been labeled inaccurate because they do not directly correspond to eye movements. We consider them as subsequent reflections on processes that are difficult to verbalize simultaneously.

To conclude, during silences, important information can be observed about users' processes and the problems they experience. Silences often occur because users have no cognitive energy left to describe what goes on in their minds. These silences should not be regarded as problematic. Not all cognitive activity, such as scanning website pages, can be easily verbalized and we should not expect participants to verbalize their thoughts during all of their actions. Therefore, in a CTA evaluation, it is sensible not to rely only on verbalizations, as precisely during silences, the most interesting observations can be made regarding users who are in trouble.

Conclusions, Limitations, and Suggestions for Future Research

In this section, we discuss the conclusions of our study within the larger context of professional communication. After the general conclusion, we will explain the "implications to theory and to practice." Further, we will discuss the "limitations" of the study, and the "suggestions for future research."

Conclusions In this study, we replicated Cooke's study [28] on the analysis of types of verbalizations, and the processes observed during silences. We chose not to replicate Cooke's research question on accuracy, as we consider an analysis of the relation between verbalizations and eye movements as an invalid conceptualization of accuracy.

The results of Cooke raised questions concerning the benefits of a CTA evaluation, as 77% of the verbalizations in her study were about easily observable *reading* and *procedure* actions. In our study, however, only 40% of the verbalizations belonged to these groups. Further, the largest group of verbalizations consisted of thoughts about the website and about the participants' own interaction with this website, such as doubts, judgments, and frustrations. So our study showed that verbalizations in a CTA study can provide information with an added value about the participants' processes and obstacles on the website.

Compared to Cooke's study, we measured a rather high percentage of silences: 27% in our study versus 16% in Cooke's study. Comparisons with Cooke on types of silences cannot be made, as she has only exploratory analyzed what occurred during silences. Silences most often occurred at moments that participants were scanning pages for information. Apparently, scanning is a quick and cognitive complex task that is difficult to verbalize. Participants seem to stop verbalizing their thoughts at that time to reduce the cognitive load [3], [13], [33], [34].

Implications to Theory: Not much research has been done yet on the types of verbalizations a CTA study generates [9], [10], [19]. Our study contributes to the theory on verbalizations in CTA studies by showing new insights into the types of thoughts that participants verbalize. Moreover, we have added to knowledge on the gaps in think-aloud protocols: the silences. It is known that participants are not able to verbalize everything that goes on in their minds [1], [2], and our study has provided new knowledge on processes that occur during silences and that seem difficult to verbalize. It might be useful to reconsider think-aloud instructions, to prevent participants from getting frustrated by prompts that remind them of verbalizing thoughts at moments when they perform actions that are difficult or even impossible to verbalize.

Implications to Practice: Practitioners who have read the study of Cooke might think that the benefits of CTA are not worth the efforts of conducting an evaluation with this method. Our study has shown, however, that a CTA study can certainly generate information that cannot easily be found with observations alone. Our study settings correspond to the usability practice, as we performed our evaluations with a diverse set of participants who frequently experienced problems

during their task performance. Professionals in usability practice can learn from our study what types of verbalizations can be expected from a think-aloud evaluation. Further, they have a better idea of the significance of silences and the relevance of the processes that can be observed during these silences. There are types of silences, for example, when people are scanning or reading, during which participants should not be forced to verbalize their thoughts. At these moments, reminders to think aloud probably only cause frustrations.

An advice for the usability practice that can be inferred from this study is that a comprehensive evaluation should combine verbalizations and observations. Our results showed that participants verbalized many thoughts that are not easily observable with other methods. But we also saw many silences, during which eye-tracking observations were necessary to shed light on users' actions and potential problems. Strictly speaking, the output of a CTA evaluation consists of the participants' verbalizations of their thoughts. However, in most evaluations in practice as well as in scientific studies, the observations of the participants' actions are also included in the analyses, such as [12]–[17].

Both information sources can independently reveal some of the user problems, but there are also cases in which an interaction between verbalization and observation leads to new insights about user problems. For example, when participants verbalize their thoughts, they regularly fail to specify the exact object their verbalization refers to, as in "This is very unclear to me." In these cases, observations of eye movements provide details on participants' actions during the verbalization, and on the object the verbalization relates to. This combination helps the evaluator make an adequate assessment of the situation and of possible problems. Another example can be found in situations where a participant verbalizes an action that is in itself neutral, but that points to a problem in combination with an observation. In our study, several participants verbalized that they would click back to the homepage and subsequently clicked on a link that did not lead them to the homepage. Without the verbalization, the evaluator would not have known that the participant had made a wrong interpretation of the link that was clicked on. And, on the other hand, without the observation of the wrong link click, the verbalization would not have clearly pointed to a problem. In this way, the combination of the two sources shows that the participant has misinterpreted something in the text. Thus, the combination of CTA and observational methods leads to the most complete overview of user problems, thanks to the added value of the interaction between both sources of information.

Limitations Our study shows large differences with the results of Cooke, on all three websites. We do not know yet, however, which adjustments in our setting influenced the outcomes of the think-aloud study in what way. We did not aim at studying the relation between specific characteristics of the research setting and types of verbalizations and silences. Therefore, we cannot draw conclusions on the exact influence of settings on the output.

Another limitation of the study is that we have only looked at one type of website, informational websites on which users look for specific information. Results might be different for other types of websites.

Suggestions for Future Research Future research on types of verbalizations should focus on the influence of research settings on the outcomes of an evaluation. For example, the influence of instructions deserves more attention. Much research has been carried out on the effect of think-aloud instructions and prompts on task processing [4]–[9], but little is known about the effects on types of verbalizations. More research is needed on the extent of prompting during task performance, and relating these prompts to the actions that participants perform at a certain moment.

Also, more attention should be paid to differences between participants in their abilities to verbalize their thoughts and the content of these thoughts. Although we used a large group of 60 participants in total, we cannot report statistical solid results because of the rather small groups of participants with different characteristics. We have done exploratory analyses on the participant characteristics that showed some differences in the types of verbalizations produced. Younger participants tended to produce less reading verbalizations than older participants. Also, higher educated participants verbalized less reading units, but they verbalized more explanations and motivations of their actions than the lower educated participants. Analyses of the percentages of silences did not show differences between age groups, or between education levels. We did not find a correlation between the task performance time and the total number of verbalizations made. This

means that a participant who spends more time on the website does not necessarily verbalize more. We observed strong differences between participants: some people who needed more time for their tasks verbalized a lot, others struggled while performing the tasks and seemed to use their cognitive energy for the task performance instead of verbalizing thoughts. Clearly, more research is needed into differences between individual participants and their verbalization characteristics. However, higher numbers of participants are needed to conduct comprehensive research on the extent to which participants differ in the way they verbalize their thoughts.

Another interesting question for future research concerns the distribution of verbalization types in the retrospective think-aloud conditions. This distribution will certainly differ from concurrent conditions, if only because of the smaller amount of reading verbalizations and the lower cognitive load. More knowledge about the differences in types of verbalizations can shed more light on the differences between concurrent and retrospective think-aloud conditions.

In our study, we have made a rather global distinction between processes during silences. It would be useful to analyze silences in CTA studies in more detail and to relate specific eye-tracking patterns to certain processes and potential problems. In several studies, such as [25], [32], and [39]–[41], this relation has been explored. However, more research is needed to deepen our knowledge on this topic. Eye-tracking data can also be used to further explore the ways in which the users' task of thinking aloud influences their behavior during task performance on websites.

REFERENCES

[1] R. E. Nisbett and T. D. Wilson, "Telling more than we can know: Verbal reports on mental processes," *Psychol. Rev.*, vol. 84, no. 3, pp. 231–259, 1977.

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

[3] T. Boren and J. Ramey, "Thinking aloud: Reconciling theory and practice," *IEEE Trans. Prof. Commun.*, vol. 43, no. 3, pp. 261–278, Sep. 2000.

[4] E. Krahmer and N. Ummelen, "Thinking about thinking aloud: A comparison of two verbal protocols for usability testing," *IEEE Trans. Prof. Commun.*, vol. 47, no. 2, pp. 105–117, Jun. 2004.

[5] M. Nørgaard and K. Hornbæk, "What do usability evaluators do in practice? An explorative study of think-aloud testing," in *Proc. 6th Conf. Design. Interact. Syst. ACM*, 2006, pp. 209–218.

[6] J. Ramey, T. Boren, E. Cuddihy, J. Dumas, Z. Guan, M. J. Van den Haak, and M. D. T. De Jong, "Does think aloud work? How do we know?," in *Proc. CHI Extended Abstracts Human Factors Comput. Syst.*, 2006, pp. 45–48.

[7] M. Hertzum, K. D. Hansen, and H. H. K. Andersen, "Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload?," *Behav. Inf. Technol.*, vol. 28, no. 2, pp. 165–181, 2009.

[8] E. L. Olmsted-Hawala, E. D. Murphy, S. Hawala, and K. T. Ashenfelter, "Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability," in *Proc. 28th Int. Conf. Human Factors Comput. Syst.*, 2010, pp. 2381–2390.

[9] T. Zhao and S. McDonald, "Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods," in *Proc. NordiCHI*, 2010, pp. 581–590.

[10] V. A. Bowers and H. L. Snyder, "Concurrent versus retrospective verbal protocol for comparing Windows usability," in *Proc. 34th Human Factors Soc. Annu. Meeting*, 1990, pp. 1270–1274.

[11] N. Ummelen and R. Neutelings, "Measuring reading behavior in policy documents: A comparison of two instruments," *IEEE Trans. Prof. Commun.*, vol. 43, no. 3, pp. 292–301, Sep. 2002.

[12] M. J. Van den Haak, M. D. T. De Jong, and P. J. Schellens, "Retrospective versus concurrent think-aloud protocols: Testing the usability of an online library catalogue," *Behav. Inf. Technol.*, vol. 22, no. 5, pp. 339–351, 2003.

[13] M. J. Van den Haak, M. D. T. De Jong, and P. J. Schellens, "Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison," *Interact. With Comput.*, vol. 16, no. 6, pp. 1153–1170, 2004.

[14] M. J. Van den Haak, M. D. T. De Jong, and P. J. Schellens, "Constructive interaction: An analysis of verbal interaction in a usability setting," *IEEE Trans. Prof. Commun.*, vol. 49, no. 4, pp. 311–324, Dec. 2006.

[15] M. J. Van den Haak, M. D. T. De Jong, and P. J. Schellens, "Evaluation of an informational Web site: Three variants of the think-aloud method compared," *Tech. Commun.*, vol. 54, no. 1, pp. 58–71, 2007.

[16] M. J. Van den Haak, M. D. T. De Jong, and P. J. Schellens, "Evaluating municipal websites: A methodological comparison of three think-aloud variants," *Government Inf. Quart.*, vol. 26, no. 1, pp. 193–202, 2009.

[17] S. Elling, L. Lentz, and M. De Jong, "Retrospective think-aloud method: Using eye movements as an extra cue for participants' verbalizations," in *Proc. 29th Int. Conf. Human Factors Comput. Syst. ACM*, 2011, pp. 1161–1170.

[18] L. Van Waes, "Thinking aloud as a method for testing the usability of websites: The influence of task variation on the evaluation of hypertext," *IEEE Trans. Prof. Commun.*, vol. 43, no. 3, pp. 279–291, Sep. 2002.

[19] L. Cooke, "Eye tracking: How it works and how it relates to usability," *Tech. Commun.*, vol. 52, no. 4, pp. 456–463, 2005.
[20] L. Cooke and E. Cuddihy, "Using eye tracking to address limitations in think-aloud protocol," in *Proc. IEEE Int. Prof. Commun. Conf.*, 2005, pp. 653–658.
[21] T. Van Gog, F. Paas, J. J. G. Van Merriënboer, and P. Witte, "Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting," *J. Exp. Psychol.: Appl.*, vol. 11, no. 4, pp. 237–244, 2005.
[22] L. J. Ball, N. Eger, R. Stevens, and J. Dodd, "Applying the PEEP method in usability testing," *Interfaces*, vol. 67, pp. 15–19, 2006.
[23] Z. Guan, S. Lee, E. Cuddihy, and J. Ramey, "The validity of the stimulated retrospective think-aloud method as measured by eye tracking," in *Proc. SIGCHI Conf. Human Factors Comput. Syst. ACM*, 2006, pp. 1253–1262.
[24] N. Eger, L. J. Ball, R. Stevens, and J. Dodd, "Cueing retrospective verbal reports in usability testing through eye-movement replay," in *Proc. 21st Brit. CHI Group Annu. Conf. HCI: People Comput. XXI: HCI . . .But Not as We Know It*, 2007, vol. 1, pp. 129–137.
[25] C. Ehmke and S. Wilson, "Identifying web usability problems from eye-tracking data," in *Proc. 21st British CHI Group Annu. Conf. HCI: People Comput. XXI . . .But Not as We Know It*, 2007, vol. 1, pp. 119–128.
[26] A. Hyrskykari, S. Ovaska, K. Räihä, P. Majaranta, and M. Lehtinen, "Gaze path stimulation in retrospective think-aloud," *J. Eye Movement Res.*, vol. 2, no. 4, pp. 1–18, 2008.
[27] T. Van Gog, L. Kester, F. Nievelstein, B. Giesbers, and F. Paas, "Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction," *Comput. Human Behav.*, vol. 25, no. 2, pp. 325–331, 2009.
[28] L. Cooke, "Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach," *IEEE Trans. Prof. Commun.*, vol. 35, no. 3, pp. 202–215, Sep. 2010.
[29] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognit. Psychol.*, vol. 8, no. 4, pp. 441–480, 1976.
[30] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychol. Bull.*, vol. 124, no. 3, pp. 372–422, 1998.
[31] J. H. Goldberg and A. M. Wichansky, "Eye tracking in usability evaluation: A practitioner's guide," in *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, J. Hyönä, R. Radach, and H. Deubel, Eds. New York: Elsevier, 2003, pp. 493–516.
[32] R. J. K. Jacob and K. S. Karn, "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises," in *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, J. Hyönä, R. Radach, and H. Deubel, Eds. New York: Elsevier, 2003, pp. 574–605.
[33] S. Monsell, "Task switching," *Trends Cognit. Sci.*, vol. 7, no. 3, pp. 134–140, 2003.
[34] M. A. Neerincx, J. Lindenberg, and S. Pemberton, "Support concepts for web navigation: A cognitive engineering approach," in *Proc. 10th Int. Conf. World Wide Web*, Hong Kong, China, 2001, pp. 119–128.
[35] S. Elling, L. Lentz, and M. De Jong, "Users' abilities to review website pages," *J. Bus. Tech. Commun.*, vol. 26, no. 2, pp. 170–200, 2012.
[36] S. Elling, L. Lentz, M. De Jong, and H. Van den Bergh, "Measuring the quality of governmental websites in a controlled versus an online setting with the "Website Evaluation Questionnaire," *Gov. Inf. Quart.*, vol. 29, no. 3, pp. 383–393, 2012.
[37] W. P. Eveland and S. Dunwoody, "Examining information processing on the World Wide Web using think aloud protocols," *Media Psychol.*, vol. 2, no. 3, pp. 219–244, 2004.
[38] A. L. Cox and M. M. Silva, "The role of mouse movements in interactive search," in *Proc. 28th Annu. Meeting Cognit. Sci. Soc.*, Vancouver, BC, Canada, 2006, pp. 1156–1161.
[39] J. H. Goldberg and X. P. Kotval, "Computer interface evaluation using eye movements: Methods and constructs," *Int. J. Ind. Ergonom.*, vol. 24, no. 6, pp. 631–645, 1999.
[40] L. Cowen, L. J. Ball, and J. Delin, "An eye-movement analysis of web-page usability," in *People Comput. XVI (Proc. HCI 2002)*. London, UK: Springer, 2002, pp. 317–335.
[41] A. Poole, L. J. Ball, and P. Philips, "In search of salience: A response time and eye movement analysis of bookmark recognition," in *People Comput. XVIII (Proc. HCI 2004)*. London, UK: Springer, 2004, pp. 363–378.

Sanne Elling is currently pursuing the Ph.D. degree at the Utrecht Institute of Linguistics (UIL-OTS), Utrecht University, Utrecht, the Netherlands. Her research project is on user-focused methods of website usability evaluation.

Leo Lentz is a professor of Document Design and Communication at Utrecht University, Utrecht, the Netherlands. His research focuses on usability and comprehension of information in different genres, including health-care information, financial documents and forms, presented in digital and hard copy formats.

Menno de Jong is a professor of communication studies at the University of Twente, Enschede, the Netherlands. He has published research articles on various methods of usability evaluation. His main research interests include the methodology of applied research techniques.