Task Variation in Usability Testing of Web Sites
An Analysis of three experimental Studies

Luuk Van Waes - UFSIA, University of Antwerp (Belgium)

In usability testing of web sites, thinking aloud is a frequently used method. A fundamental discussion, however, about the relation between the use of different variants of thinking aloud and the evaluation goals for this specific medium is still lacking. To lay a foundation for this discussion I analyzed the results of three usability studies in which different thinking aloud tasks were used: a simple searching task, an application task and a prediction task. In the task setting the profile of the web surfer, the communication goal of the web site and other quality aspects are taken into account. The qualitative analysis of these studies shows that the task variation has some influence on the results of usability testing and that, consequently, tasks should be matched with the evaluation goals put forward. .

Thinking aloud is a widely used but controversial research method that can be applied for studying problem solving as well as for evaluating purposes. In order to study problem-solving processes researchers have developed different procedures for conducting thinking aloud research. However, in the literature about the thinking aloud method and in the analysis of thinking aloud protocols hardly any attention has been given to the influence of the task on the quality and the characteristics of the test results. Although Breuker [1] already mentioned the influence of the expressibility of the task, the characteristics of the main task, the verbal capacities of the participants and their familiarity with the task, only few researchers have continued along those lines.

A substantial contribution to thinking aloud research has been made in the field of writing research. The discussion about the influence of the task, however, has mainly been limited to a methodological debate on the *reactivity* of concurrent thinking aloud protocols on the main (writing) task. The main question was whether and how the writers' cognitive processes are disrupted by the fact that they are writing and thinking aloud at the same time. [2-7]

This article focuses on a different aspect of task setting. I would like to investigate how the *thinking aloud task* in usability testing of web sites influences the kind of results that can be obtained from the analysis of the thinking aloud protocols. From a usability point of view: which variants of thinking aloud tasks contribute best to realize specific evaluation goals and quality aspects? The research focuses on the searching behavior of web users and on detecting navigation barriers they encounter in these web sites.

Task types for thinking aloud protocols

In digital environments thinking aloud is widely used for usability testing. In most of the research on web site usability, tests are set up in which a user is given a set of realistic tasks and asked to perform them using some version of a web site while thinking aloud. Standard statistics, such as task completion rates and times, are commonly tracked, along with usability issues derived from the analysis of the protocols.

Often this method is used in combination with other evaluation techniques. Gray [8], for instance, used protocol analyses in combination with user drawings of the organization of information on web sites to investigate the problem of 'getting lost' in hypertext navigation. Tullis et al. [9] combine thinking aloud tests with evaluation techniques like *'timed click tests'* (Where would users click on this screen shot to perform a certain task?), *'eye tracking tests'* (Where do users actually look on the screen?) or *'attention tests'* (What captures the users' attention on a web page?).

The description of several usability tests using thinking aloud protocols also reveals that the tasks used are somewhat different. However, the characteristics of the task itself are seldom explicitly discussed. Spool [10-11] is one of the few researchers who focuses on the nature of different tasks. He describes four types of tasks: simple facts (e.g. *'Can you get a Honda Accord for under $15,000?*), comparisons of facts (*'Which is cheaper to fly to, Nevada or England?'*), judgment (*'Would you like to go on a day trip to Hampton Court?'*) and comparison of judgment (*'In your opinion, what's the best new convertible for under $20,000?')*. On a more detailed level, he describes how to construct usability tasks, stressing the importance of the realistic character of the task and the user's commitment to the task. One of the strategies he describes, for instance, is that he interviews visitors of antique markets about the things they are looking for in order to construct tasks to test an on line auction site of antiques. It is obvious that those field-derived tasks are more realistic than desk-constructed tasks of non-experts. Basically, however, the tasks used are all simple searching tasks that focus on information retrieval.

Although a lot of researchers are aware that task type is an important factor in the design of adequate web usability tests, a fundamental discussion about the relation between different variants and evaluation goals is still lacking. In this article I would like to initiate this discussion with a qualitative interpretation of the results of three studies in which different task types were created to detect navigation barriers in web sites. The focus in the discussion is always on matching the task used in the thinking aloud research with the quality goals put forward in the usability test. But before dealing with my own research, I will give a short overview of the most important aspects of web site usability and I will briefly comment on some of them.

Aspects of web site usability

Schriver [12] classified evaluating methods into text focused methods (an evaluation by the writer), expert-judgement methods (an evaluation by professionals with expert knowledge of the subject, the target audience, or texts) and user or reader focused methods (evaluation by the target audience of the text). In this article I will concentrate on the formative

evaluation of web sites from the users' point of view[1]. In this kind of research several usability aspects should be taken into account when setting up a design. For instance:

- the user's aim and related surfing behavior;
- the quality objectives of usability test;
- the possibility of measuring (quantitative) efficiency (e.g. time needed);
- the possibility of measuring (quantitative) effectiveness (e.g. number of correct answers) ;
- the focus on certain characteristics of the web site itself (text and graphics) and its goals (informative, persuasive, divertive or instructive);
- the focus on the different components of the web site (home page, frames and navigation tools, forms, specific content branches, java scripts).

To set up a valid framework for the discussion of the strengths and weaknesses of the different task types (cf. discussion section), I will first discuss two - more complex - aspects: surfing behavior and quality objectives.

Surfing behavior

Web users fall into different traditional categories, according to their sex, age, interest, background, web experience etc. (See also Hackos & Redish [14] about user and task analysis). However, Muylle et al. [15] clearly show that the user's search behavior also is a very important factor to take into account when evaluating web sites. In the data of their exploratory multiple-case study with consumers and business people they discovered five behavioral search scenarios based on how goal-oriented and specific the search behavior was:

- exploratory surfing: low purposiveness search objective with the surfer looking for something new, something cool and skimming lists of hyperlinks (low specificity);
- window surfing: the surfer is not looking for any specific information, but the attention is attracted by fortuitously discovered content;
- evolved surfing: the focus is on a few particular items within the boundaries of the triggered search content ;
- bounded navigation: the search boundaries are determined a priori and a broad range of search goals is addressed;
- targeted navigation: the surfer searches determined and specific content (high-purposiveness and high specificity).

This description shows that each of the surfing profiles is related to a different task setting. Consequently, when choosing an evaluation method it is important to take into account that these different searching scenarios exist and that they may require a different task approach. Conversely, when formulating a usability task we should be aware that we 'feed' a certain surfing profile, while suppressing others.

Quality objectives

'The more a site helps people find the information they are looking for, the more usable it is.' [10:2]. It is beyond doubt that information retrieval is an important goal in web site usability, but it is certainly not the only objective in usability testing. A very useful

classification of quality objectives can be found in De Jong & Schellens[13] and De Jong & Heuvelman [16]. They break down the overall concept of effectiveness into a series of conditions for effectiveness and identify the following quality objectives:

- contact: how is the (information on the) web site brought to the potential readers' attention?
- selection: how does the target group select the information within the site?
- comprehension: does the reader understand the information on the site correctly?
- application: is the reader able to apply the information in a realistic setting (especially for instructive texts)?
- acceptance: is the presentation of the information acceptable to the readers?
- appreciation: are the style, the tone, the register, the design appreciated by the readers?
- relevance and completeness: are the most important topics in the site relevant for the intended reader?

It is impossible to address the full spectrum of quality objectives with a single task or a one dimensional task setting. Therefore, when developing ideas for thinking aloud tasks, it is important to strike the right balance between the type of texts and themes that are focused on in the evaluation and the other aspects of web site usability mentioned earlier in this section.

Three studies and three task types

In this article I will reflect on the results of three different usability studies. In each of these studies a different variant of thinking aloud is used to evaluate different (parts of) web sites.

- Study 1: SIMPLE SEARCHING TASK
  In a traditional usability test design the participants had to find specific information (simple facts in Spool's terminology [10]) on two web sites of Belgian banks. The objective was to find answers to a number of specific questions. During their search they had to produce a concurrent thinking aloud protocol. The central aim of the study was to detect navigation barriers in the process of selecting information and to gather information about the acceptance of the information.
- Study 2: APPLICATION TASK
  The participants had to read and reconstruct a route description presented on selected web sites while producing a thinking aloud protocol. Not the information retrieval (study 1), but the applicability of the information presented was the focus of this study.
- Study 3: PREDICTION TASK
  The participants had to predict what information they would find behind the links on the home page of two financial and two insurance companies without clicking. This study focused on the comprehension of the links presented on the home page (classification model, terminology, combination of text and graphics etc.) and on the mental models the users develop.

I will now briefly describe the design and the results of each of the studies. To illustrate the potential strengths and weaknesses of every task, I have selected five categories of navigation problems for each of the studies that were representative of the quality

objectives the users mainly focused on. Each of these categories will be illustrated with examples from the thinking aloud protocols. In the last section I will discuss some additional usability aspects when comparing the different research tasks I used in the thinking aloud sessions. Of course, we should take into account that in the three studies different web sites have been evaluated.

Study 1: Simple searching tasks and thinking aloud to compare on-line and off-line searching behavior

The main objective of this study[2] was to gain a better insight into the kinds of thinking aloud tasks that enable us to describe the general searching and navigation behavior of people looking for detailed information on a web site. Using a traditional usability design, we asked our participants (n=12) to look up information about specific topics on-line and off-line. For the on-line mode we chose the web sites of two Belgian banks (Axion and BBL) offering special services for young persons; for the off-line mode we used the brochures of the same banks offering comparable information. We selected these banks on the basis of an evaluation of 7 web sites using 42 criteria related to content, navigation, graphics, frames etc. (e.g. amount of information, language options, graphic hyperlinks, navigation frames, interactivity, search function).

To collect data about the searching behavior of the participants in the on-line mode we used an on-line camcorder (Microsoft Camcorder™). This program enabled us to make a timed recording of all the mouse movements and browser actions and to record a simultaneous thinking-aloud protocol without disturbing the navigation process. To avoid differences in waiting time due to traffic jams on the Internet, we downloaded the two web sites (Grab a Site™) and put the sites on a local server. In the off-line mode we limited the observation to audio recorded thinking aloud protocols. The study was set up in a Latin-square design. At the end of the sessions the participants were interviewed about their experiences.

Participants
Twelve persons participated in this small-scale study. All of them were students (18-24 years old). The students (6 male - 6 female) had different backgrounds and they all had a bank account at a bank that was not selected for the study. Half of them were experienced Web surfers (about 10 hours a week); the other half had had very limited experience with the Internet.

Assignment
The thinking aloud sessions began with a test assignment. Then the observer gave the participants the tasks one by one on paper. The observer asked them to read the task aloud and regularly stimulated them to keep on talking. The tasks were very specific and were divided over the different content components of the web site , e.g. what's the maximum amount of money you can borrow from the bank when you have a Y-account?; which credit cards are free when you have a Y-account?

Analysis
The data of the thinking aloud protocols were analyzed from three different perspectives:
- accuracy: was the answer correct and detailed enough?
- time: how much time did the participant need to find the answer?
- navigation strategy (routing): which navigation tools did the participant use?
- navigation problems: what were the navigation barriers the participant came across?

Results
For a detailed time analysis and the results of the accuracy test, I refer to an earlier publication [17] because these analyses are beyond the scope of this article. I can mention briefly that on average the participants in the on-line mode (Internet) needed more than twice the time to complete the tasks than in the off-line mode. The results of the accuracy analysis show that the participants answered significantly more questions correctly on-line than off-line. The analyses of the thinking aloud protocols also showed that the navigation tools offered by the two sites seemed to give rise to different searching strategies (cf. infra).

This paper focuses on detecting and diagnosing the navigation barriers participants encountered when carrying out simple searching tasks. As it is, the searching behavior in the on-line mode proved much more divergent than in the off-line mode. There was no consistent preference for certain navigation tools on-line. Probably depending on their searching profile, some participants preferred the graphical navigation buttons, while others only used the textual hyperlinks, either in the frame or in the body text; some participants always used the browser interface to go back to the previous page, while others always looked for a home- or back-button in the body of the text.

To illustrate the possibilities of this research design, I will list some categories of navigation problems and illustrate them with examples from the thinking aloud protocols which are typical of the kind of navigation problems that can be detected with this type of simple searching tasks. I have also tried to link them to a specific quality goal, which is mentioned between brackets.

- *Non-explanatory titles* (selection and appreciation problems)
  One of the qualities of hyperlinks which Schriver [18: 391] describes is that the name of a link should give a good indication of the content behind it. Readers should be able to forecast accurately what the content of the information is behind the link. The protocols give us some clear examples of hyperlinks that do not seem to describe the content clearly enough.
  Examples: One of the highlighted links on a web site was 'leen een been' (translation: 'borrow a leg'). The nonsensical collocation was probably meant for rhyming purposes only. The information behind this link dealt with possibilities and restrictions of borrowing money from the bank. Another link that caused confusion was 'Become as rich as the sea is deep, as an elephant is fat or as a traffic jam to the sea side in summer is long'. This kind of phrasing not only puzzles the readers because they are not able to predict the content behind the link, but also irritates them because of its lack of functionality.

- *Lack of visual consistency and non-exploratory graphics* (selection problem)
  In one of the sites, the participants had to switch from a framed interface to a frameless screen with a dial in the upper left corner (figure 1). The rest of the screen is white. If you click on the numbers of the dial, a text appears on a certain topic. The inconsistency - and the implicitness - of the interface design puzzled most of the participants. Moreover, the different topics only appear after you have clicked a number on the dial. No direct topic indication is used on the dial.

*** Figure 1 somewhere here ***

- *Non-distinguished link types* (selection and comprehension problems)
  Links to definitions of financial terminology (deposit account, options etc.) are often represented in the same way as links to more specific information about a topic. Also, more functional links, such as the 'help'-link, are not represented differently from the topical links. A lot of participants were also confused by this 'help'-link for another reason. It often caused navigation problems because the function was phrased as if it represented a site-specific help function as you can find in most software programs. Therefore they did not expect a within-site search behind the link.

- *Spelling and grammar mistakes* (appreciation problem)
  A simple spelling error as in the word 'doc***i***mentatiecentrum' ('doc***i***mentation center') was noticed by all participants. It distracted them and some even expressed their annoyance.

- *Screen exceeding texts* (selection problem)
  In some of the screens the text - and some of the links - exceeds the 15-inch screen. Reading the complete text requires scrolling and this often caused a navigation barrier for the participants in the study. Some of them did not notice the scroll bar at the right that indicated that the text exceeded the visible zone on the computer screen.

Conclusion

The combination of simple searching tasks for thinking aloud with an on-line registration of the searching process proves to be a valuable observation method to collect data about general navigation strategies in evolved, bounded or targeted navigation. Selection and appreciation problems are the main quality objectives that are evaluated. Barriers that were connected to other quality objectives were hardly detected. Measuring efficiency and effectiveness can easily be combined with this type of task setting.

Study 2: Application tasks and thinking aloud to evaluate the applicability of route descriptions on the Internet

In this second study[3] our evaluation focus was on the applicability of instructive information on web sites. With this study we tried to fine-tune the evaluating techniques by applying them to specific sub-genres of on-line texts. Different parts of web sites reveal different sender aims and can focus on persuasive, informative or instructive aspects. As an example of instructive texts we chose route descriptions as they are part of many web sites. A route description aims to tell a potential visitor where the company or institution is physically located and how to reach that location [19-20]. We asked our participants (n=10) to think aloud while reading and interpreting the route description and to reconstruct the description afterwards without using the on-line information. We chose two web sites: the site of the University of Amsterdam (UvA) and the site of Media Plaza in Utrecht, both located in the Netherlands. We selected these sites on the basis of a descriptive analysis of route descriptions on the Internet[4]. Both sites were characterized by a high degree of interactivity to suit many users' specific needs (different means of transport, points of departure, scales and scopes of the maps, links to other sites, such as one with options of public transport etc.). The presence of graphics or other multimedia elements was another selection criterion.

To collect the participants' protocols we used the on-line Microsoft Camcorder™ again (cf. first study) and in addition a normal camcorder to record the route reconstruction afterwards. To avoid order-effects, the sequence in which the sites were shown to the participants was counterbalanced.

Participants

Ten persons (5 male - 5 female) participated in the study. They were between 18 and 26 years old. The participants had different academic backgrounds and all had a driver's license. None of them were familiar with the direct neighborhood of the locations selected. Half of them were experienced web surfers (about 10 hours a week); the other half had had very limited experience with the Internet.

Assignment

The thinking aloud sessions began with a short test assignment to give the participants the opportunity to familiarize themselves with the thinking aloud technique. Then the observer gave the participants the main task. She instructed them to read the route description while thinking aloud. At the end of the session the participants had to be able to reconstruct the optimal route from the KUB-campus (location of the study) to the location presented on the web site. They were allowed to take notes and use road and city maps. During the reconstruction afterwards they were allowed to use the additional material, but not the web site. The observer regularly stimulated the participants during the study to keep on thinking aloud. After having finished the double assignment, the participants were interviewed about their experiences and they were asked to to evaluate the on-line route descriptions on different semantic scales.

Analysis

The data of the thinking aloud protocols were analyzed from three different perspectives:
- accuracy: was the application correct and detailed enough?
- time: how much time did the participant need to find and reconstruct the route?
- navigation strategy: which navigation tools did the participant use?
- navigation problems: what were the navigation barriers the participant came across?

Results
Similar to the first study, I will focus on the results of the navigation strategies and present a limited set of qualitative results, complete with examples from the thinking aloud protocols. The selection of these results is meant to be representative for the kind of navigation problems that can be detected and diagnosed in an application task for thinking aloud (instructive texts on web sites).

- *Missing points of reference* (application problems)
  On the UvA-web site graphic maps of the neighborhood of the campus are given. The users can choose between three different maps with different scales linked hierarchically upwards (from less detail to more detail): a map of the campus itself, the immediate neighborhood and the broader region. The problem we came across was that most participants who started at the second level (figure 2) were not able to find adequate reference points to orientate themselves on this map.

*** Figure 2 somewhere here ***

  The approach roads were only indicated on the map with the smallest scale. Although there was a textual link at the bottom of the page, most participants did not see this link because it was presented in the scrolling zone (cf. position of the vertical scroll bar in figure 2). Moreover, the map itself included only few landmarks or street names, which made it difficult for the participants to construct an adequate cognitive map or to transfer the map to the city map on paper.

  It is interesting to mention that in another study, in which we asked the participants to surf on this site without any tasks (exploratory or window surfing), this part of the route description was evaluated very positively and no application problems were reported.

- *Transfer of textual information* (application and appreciation problems)
  Another application problem we came across was that some participants had difficulties in transferring the on-line textual information to the graphic representation of the off-line maps. This transfer problem was revealed differently in both web sites. On the UvA-site, for instance, the textual and the graphic sets of information were presented on separate pages and consequently could not be displayed simultaneously. Most participants tried to transfer the textual information to the off-line maps but experienced trouble matching the information. An example of part of the text:

  - Take the Ring road (A10) direction 'Center, Hilversum, Amersfoort'

- Exit 'Center, Duivendrecht, ...' (S112)
- At the traffic lights turn right 'Center'
- At the roundabout turn right, direction 'Center': Wibautstraat
- After about 1000 m, behind the 'Weesper Arcade' building turn to the right [...]

At first sight this looks like a very clear and detailed route description: clearly delimited, instructive steps, a lot of secondary information at different levels (technical indications, street names, references to buildings) etc. For some participants, however, step 1 of the route description already created a problem. They did not know whether to follow the ring road to the left or right because the direction information was not indicated on the map. A further look at the map (identifying the cities indicated) was a solution for most of the participants. A tougher problem was created in step 3. Although information about traffic lights is often a useful indication when driving, this kind of instructive information is impossible to transfer to a traditional map. Nor is the extra secondary information ('Center') useful. Those who tried to solve the problem by using the electronic maps did not get any further. The traffic lights or the roundabout were not indicated either and the street names were not legible on the screen. The only solution was to look for the 'Wibautstraat' (indication in step 4) in the index of the paper map and to reconstruct the route description backwards from there.

- *Transfer of on-line graphic information to off-line graphic information* (application problems)
  As mentioned above, most participants preferred a traditional map when actually driving to the location. So they tried to transfer the graphic information from the screen to paper. To illustrate the problems they experienced in this process, I would like to refer to the Media Plaza web site. In this web site an interactive route planner makes it possible to enter the postal code of the starting point to get a graphic representation of the route. This map has several zoom-in possibilities, but the further one zooms in, the more abstract (and deformed) the maps become. Most participants faced problems transferring the abstract road patterns to the paper map because they were presented so differently. In addition, most people did not notice the textual description under the map on-screen because this information was 'hidden' in the scrollable zone, separated by a commercial banner. Most of the participants thought this banner was the end of the page and did not consider scrolling down.

- *Missing information* (completeness problems)
  A very elementary piece of information in a route description is the identification of the destination (point information). In the UvA web site, for instance, the description starts with the address of the main faculty building (and telephone number). The Media Plaza site, however, has no such information in the route description This was explicitly mentioned in the protocols of those participants who wanted to find the exact location of the destination on a street map or the telephone number in case of traffic jams. Again this shows a clear link between the thinking aloud task and the searching behavior.
  More illustrations of this category of problems include the forecasts that were expressed in the protocols. Some participants, for instance, were looking for a picture of the

building of the destination in order to be able to identify the destination more easily when arriving in the neighborhood. This information was offered by neither of the sites.

Conclusion
The qualitative analyses of the thinking aloud protocols supports the hypothesis that an application task is a valid way of evaluating the quality of the instructive components of the web site (i.c. route descriptions). Especially application problems were detected, sometimes in connection with appreciation issues and missing information (completeness). The dominant navigation strategy for this type of task can be identified as targeted navigation. This kind of application tasks also makes it possible to evaluate the effectiveness and the efficiency of the instructions.

Study 3: Prediction tasks and thinking aloud to evaluate the clarity of hyperlinks on home pages

The main objective of the third study was to evaluate the use of hyperlinks on the home page of web sites. With this in mind I worked out a third variant of the thinking aloud method. Inspired by the experimental design of Van der Vlist described in Schriver [18], I concentrated on the forecasts people formulated about the links presented on the home page. In the study the participants (n=20) received two tasks. The first was a general prediction task in which they had to express systematically what information they expected to find behind every link on the home pages presented to them. The second prediction task was different in that the participants had to predict which link contained the information about a specific theme. Before clicking they had to tell which link they would activate on the home page to find more information about a particular theme. In this way I have tried to evaluate whether the characteristics of the links helped users to anticipate their content accurately.

To collect data about the prediction patterns and their cognitive basis, I used an on-line camcorder (HyperCam™, a Hyperionic's on-line camera). This program creates compact AVI-files by capturing an adjustable number of screens per second together with a recording of the thinking aloud protocol produced by the participants. With this product a timed replay is possible, which makes it easier to analyze the protocols than with the compressed MS Camcorder files.

Web sites and links
I selected four home pages for this study in order to set up a Latin-square design. Both the thinking aloud tasks and the sites were counterbalanced. The sites were chosen because of the link types used on the home page.

Baron & Cary [21] give a short review of some taxonomies of link types that have been used. In their study they also develop their own taxonomy of links and provide an experimental basis for it. They differentiate between organizational links and content-based links. *Organizational* links describe the surface structure of the document (e.g.

previous or next page). *Content-based* links, on the other hand, deal with the meaning of the text. Three categories of relationships reflect these associations: semantic (e.g. a 'part of' link to organize the physical arrangement of the text), rhetorical (e.g. an 'illustration' link to support the task or to achieve a learning goal), and pragmatic (e.g. a 'usage' link to express the relation between the text and its use).

Hackos and Stevens [22] for their part focus their classification exclusively on the functionality of links and distinguish different functions of hyperlinks:

- to address different audiences
- to point to related information
- to indicate cross-references
- to define browsing paths
- to refer to common topics
- to control topic size and appearance
- to display or zoom in on graphics or other multimedia files

In selecting the sites I added a number of extra dimensions:
- phrasing: word, phrase, sentence
- appearance: text, icon, (animated) graphic
- physical position: in a frame or in the body, horizontal or vertical, at the top or at the bottom of the screen, left or right, top level or sub level in the enumeration
- perspective: topic, sender or user
- function (cf. list above)
- explicitness: restricted to the link, extended with the immediate context of the link, extended with a java script

On the basis of this classification I selected the following four sites, which had a wide variety of links on their home page: the sites of two Belgian banks, GBank and BBL, and of two Belgian insurance companies Royal Belge and OMOB.

Participants
Twenty persons participated in the study (12 male - 8 female; 22-54 years old). They had different academic backgrounds and varying degrees of Internet experience.

Assignments
Two different prediction tasks were used in this study. The participants were either instructed to express what information they expected to find behind every link on the home pages presented to them or to tell which link they would activate on the home page to find more information about 15 specific themes presented to them (e.g. annual report of the bank, insurance for home personnel). As for the latter tasks, participants were allowed to click on two links after they had finished the complete assignment to evaluate and/or correct their hypotheses. For the prediction tasks the participants were asked to move the cursor to the link they were talking about. Possible java scripts could be activated in this way too.

Analysis

The data of the thinking aloud protocols were analyzed from two different perspectives:

- navigation strategies and (cognitive) problems in predicting the information behind the link;
- accuracy: was the answer correct? (second prediction task)

In this article I will limit the result section to the navigation strategies.

Results

As in the previous studies I would like to give examples of five dominant navigation problems I came across when analyzing the thinking aloud protocols.

- *Perspective of links* (comprehension and selection problems)
  As mentioned above, links can be defined from different perspectives (topic, sender or user). The expectation protocols clearly showed that mixing perspectives often confused the participants. Especially the general prediction task revealed this kind of confusion caused by a change in perspective in the classification of the links. In the BBL web site, for instance, four links were presented one below the other (left frame): private, younger than 26, professional and BBL. The first three links were intended to refer to the target groups, the fourth was topic oriented (annual report, structure of the company etc.). However, some participants thought that the BBL link referred to information for employees of the bank, a fourth target group. This interpretation was strengthened by the presence of another BBL link (logo) at the bottom of the same frame, although this referred to information about the merged bank.
  A comparable problem arose in the OMOB web site. A horizontal frame presented both links for products and services. Hardly any of the participants' forecasts in both task modes differentiated accurately between the two links. Especially in the specific prediction task, most participants preferred a topic oriented link structure to a target group oriented structure.

- *Visibility of links* (contact problems)
  One of the most interesting aspects of the analysis of the protocols was the set of concepts used by the different participants in orienting themselves in a home page. Some participants just used a few seconds to look at the global structure of the home page and immediately started with the task, while others systematically scanned the different parts of the home page often in a completely different way. Neither of these strategies, however, seemed a perfect remedy for not skipping certain links. Especially in the theme oriented prediction task problems with the physical location of links arose. In the OMOB home page, for instance, two rows of five items are presented in a horizontal frame on top of the page. An eleventh item, 'news', is presented in the same frame, but in another color (red) and in another direction (vertically). Many participants did not notice this link or noticed it very late.
  A comparable problem was identified in the BBL home page, in which a vertical frame on the left and a horizontal frame at the bottom of the page were combined. Those participants who started with the frame on the left noticed the other frame only after a long time. For most of these participants the body of the page with a double column structure and a flashing theme box seemed to have been a serious distraction.

- *Interrelated predictions* (comprehension and selection problems)
  Because links are often presented in groups, most participants tried to base their predictions on the thematic relation between the links and did not rely on the link in isolation. This cognitive strategy revealed different problems in the presentation of links. On the Royal
  Belge home page for instance five links were grouped under the heading 'SAVE': taxability & pension, strategic savings, long term savings, save during a medium long period, child savings. Many participants were confused by the different characteristics of savings used in this classification (system, technique, target (group), period) and variation in grammatical structure, especially during the general prediction task.
  On the home page of the GBank the links Hotnews, Infostand and bank transactions were presented together. Most participants experienced this grouping as very confusing and for different reasons (unity of classification, internal and external overlap). This is illustrated by the wide range of forecasts for the Hotnews link: press releases, general financial news of the day, important political news, up-to-date stock market prices etc.

- *Use of multiple advanced organizers* (contact)
  Adding paraphrased and/or more specific information to a link is an important guideline in most web site guides. Making a link more explicit makes it easier for the reader to make an exact prediction. This goal can be achieved by different verbal and graphic techniques. A very popular technique used nowadays is the use of java scripts with extra, graphical and/or verbal information, which are activated by moving the mouse over a link. This technique is applied on the home page of GBank (figure 3). Because the graphic and the verbal scripts to the left and the right of the link were activated simultaneously, many participants did not notice the explanatory, verbal information in the text box and only noticed the changing (less explanatory) graphic information on the right. Only the more experienced users were able to benefit from the additional information.

*** Figure 3 somewhere here ***

- *Hierarchy of links and thematic inclusion* (comprehension problems)
  A last example of how the prediction tasks influence the evaluation deals with the structural hierarchy of links and their thematic inclusion. On the Royal Belge home page, for instance, the different products and services are presented in an attractive, circular way. In addition, two links are presented in a colored horizontal frame at the top of the page: 'products and services' and 'home page of AXA Royale Belge'. The latter does not refer to the home page on the screen (auto-reference) as most of the participants predicted, but to the merged company (AXA) of Royal Belge. In the same way, the 'products and services' link does not refer to the products and services listed on the home page shown (presented in a list structure as most of the participants predicted), but to a page in which the different insurance companies of the group are presented.

This is obviously a problem with the structural hierarchy of links, which is often too implicit.
Repeating the same link in different locations on the home page and sometimes with varying representations seems to cause similar problems. On the OMOB page, for instance, the link 'your insurance policy' is presented twice and is also included in the 'product' link. Many participants tried to adapt their forecasts about these links during the general prediction sessions because they thought that thematic inclusion or overlap was not very likely.

Conclusion
The typical categories of problems discussed above indicate that comprehension, selection and contact are the main quality aspects that can be evaluated by prediction tasks. The analysis also reveals that the more general tasks complement the more specific tasks when thinking aloud. Giving the participants the opportunity to verify their predictions by limited clicking was also a useful supplement. It clarified certain cognitive processes of the searching behavior and revealed the way in which mental models about the site structure are constructed, verified and corrected. Measuring the effectiviness of the site was limited, and the evaluation was mainly restricted to the quality of the hyperlinks and other navigation tools on the homepage (classification, terminology, java support etc.).

Discussion
The combination of thinking aloud protocols with on-line registration of the navigation process has proved to be a valuable observation method to collect data about various usability aspects, navigation strategies and cognitive aspects of the searching process. The method could also be recommended for pretests and usability research.

This article has focused on three variants of task types for the thinking aloud method. In the studies we asked the participants to think aloud while looking for information on a web site from different perspectives. These perspectives were realized in three task types: simple searching tasks, applications tasks and prediction tasks. The qualitative data from these three studies showed that the characteristics of the thinking aloud task influence the outcome of the research. Depending on the task, the participants focused on different quality aspects and expressed different kinds of usability problems, navigation problems and strategies. Table I is an attempt to summarize the preliminary findings from the previous sections. In the columns the different types of tasks are compared on the basis of the different usability aspects described in the second part of this article.

Table I. Overview of quality aspects evaluated by different task types

	Simple searching tasks (study 1)	Application tasks (study 2)	Prediction tasks (study 3)
User's aim and	evolved	targeted	bounded

	Simple searching tasks (study 1)	Application tasks (study 2)	Prediction tasks (study 3)
surfing behavior	bounded targeted		targeted
Quality objective	selection appreciation	application completeness (appreciation)	comprehension selection contact
Communication goals	informative (instructive)	instructive	informative
Components of the web site	homepage selected components navigation tools	instructive components	hyperlinks and navigation tools on homepage
Possibilities for efficiency measurements (e.g. time needed)	many	some	limited
Possibilities for effectiveness measurements (e.g. correctness of results)	many	some	limited

As table I shows, only a limited number of surfing types are addressed by the three types of tasks described. An evaluation of web site usability for exploratory surfing, for instance, probably requires a completely different task in which users are challenged to freely explore a web site. The same holds for the quality objectives and communication goals that can be covered with the three types of tasks. The second study illustrates, for instance, that focusing on one communication goal of a web site (i.c. instruction) requires a specific thinking aloud task and impacts on other quality aspects. This seems to suggest that divertive, persuasive and informative communication goals require a targeted approach too.

The choice of task also limits the scope of the evaluation. In more positive terms it enables us to focus on certain components of a web site more thoroughly. The prediction task, for instance, proved to be a very good way of detecting navigation barriers on the homepage in relation to specific characteristics of hyperlinks. Finally, the table also shows that certain tasks allow for measuring controlled efficiency and effectiveness, which may be important when comparing different options.

Based on the tests described in this article, it is clear that additional research is warranted to further explicate how think-aloud usability tasks can reveal the different quality aspects of web sites. Especially experimental research in which the above task types for thinking aloud are compared in a well defined web site usability context is needed. This kind of research should give us a better insight into the influence of think aloud tasks on usability testing in general and should give an answer to more specific questions about quality aspects raised in this qualitative study. One of the aspects in this study that requires further attention, for instance, is the ype of information in different parts of web sites. In

the different studies in this article different web sites have been used. As a result, the comparison of task types may have been partially distorted by certain characteristics of the different web sites involved in the study. Experimental research in which task types are varied in the usability testing of one web site is really needed to give a more decisive answer about these aspects.

At this moment only few methods for usability testing of web sites are validated adequately. Research along the experimental lines explained above would contribute to a well-founded method for usability testing. We really need further research to create a basis for web usability that is built on more than just practical experience and (plausible) assumptions. Thinking aloud has had a long tradition in different research disciplines, and really deserves to be ranked high in future research.

Notes

Acknowledgment
My thanks to my colleagues, Geert Jacobs and Gilberte Lenaerts, for their patient readings and helpful suggestions.

Literature

[1]Breuker, J.A., J.J. Elshout, M.W. van Someren & B.J. Wielinga (1986). Hardop denken en protokolanalyse. [Think aloud and protocol analysis.] *Tijdschrift voor onderwijsresearch*, 11, 241-254.

[2]Ericsson, K.A., & Simon, H.A. (1993 2nd). Protocol Analysis; Verbal Reports as Data. MIT Press: Cambridge, Massachusetts, London.

[3]Ransdell, S.E. (1995). Generating thinking-aloud protocols: Impact on the narrative writing of college students. *American Journal of Psychology*, 108/1, 89-98.

[4]Russo, J.E., Johnson, E.J. & Stephens, D.L. (1989). The validity of verbal protocols. *Memory & Cognition*, 17, 759-769.

[5]Smagorinsky, P. (1989). The Reliability and Validity of Protocol Analysis. *Written Communication,* 6, 463-479.

[6]Smagorinsky, P. ed. (1994). *Speaking about Writing. Reflections on Research Methodology.* Thousand Oaks / London / New Dehli: Sage Publications.

[7]Janssen, D. & L. van Waes 1996). Effects of thinking aloud on writing processes. In: C.M. Levy & S. Ransdell (eds.), *The Science of Writing. Theories, methods, individual differences, and applications.* New Jersey: Lawrence Erlbaum, p. 233-250.

[8]Gray, S.H. (1990). Using protocol analyses and drawings to study mental model construction during hypertext navigation. *International Journal of Human-Computer Interaction,* 2/4, 359-377.

[9] Tullis, T.S., E.J. Dixon & H.M. Hersh (2000). A "Bag of tricks" for Web Usability. G. Szwillus & T. Turner (eds.), *CHI 2000: The Future is here (extended abstracts).* New York: ACM SIGCHI, p. 306.
[10] Spool, J. et al. (1997). *Web Site Usability: a Designer's Guide*. North Andover, User Interface Engineering.
[11] Spool, J. et al. (2000). *Measuring Web Site Usability.* Workshop at CHI 2000, The Hague.
[12] Schriver, K. (1989). Evaluating text quality: The continuum from text-focused to reader-focused methods. *IEEE Transactions in Professional Communication*, 32, 238-255.
[13] De Jong, M. & P.J. Schellens (1997). Reader-Focused Text Evaluation. An overview of Goals and Methods. *Journal of Business and Technical Communication*, 11/4, 402-432.
[14] Hackos, J.T. & J.C. Redish (1998). *User and Task Analysis for Interface Design.* New York: John Wiley & Sons.
[15] Muylle, S., R. Moenaert & M. Despontin (1999). A grounded theory of World Wide Web search behaviour. *Journal of Marketing Communications*, 5, 143-155.
[16] De Jong, M. & A. Heuvelman (1999). De formatieve evaluatie van voorlichtingssites op het World Wide Web: een inventarisatie van benaderingen. [Formative evaluation of public information on the WWW: a survey of approaches] Van Ruler, A.A., P.J.M.C. Schellens, O. Scholten et al., *Jaarboek Onderzoek Communicatiemanagement,* Alphen a/d Rijn: Samsom.
[17] Schriver, K.A. (1997). *Dynamics in Document Design: Creating Text for Readers*. New York: John Wiley & Sons.
[18] Van Waes, L. (1998). Evaluating on-line and off-line searching behavior: using thinking-aloud protocols to detect navigation barriers. C. Hughes et al. red., *Scaling the heights: the future of information technology*. Québec: ACM Sigdoc, p. 180-186.
[19] Van Waes, L., Vanherreweghe, I. & Verhetsel, A. (1997), Routebeschrijvingen: Een samenspel tussen talige en cartografische instructie. [Route descriptions: A concerted action between linguistic and graphic inforamtion] H. van den Bergh, D. Janssen, N. Bertens & M. Damen (red.), *Taalgebruik ontrafeld.* Dordrecht: Foris Publications, p. 401-414.
[20] Van Waes, L., Verhetsel, A. & Vanherreweghe, I. (1998), Controle-informatie in routebeschrijvingen. [Control information in route descriptions] *Taalbeheersing*, 20/2, 141-154.
[21] Baron, L. & T. Cary (1996). Labeled, Typed Links as Cues when Reading Hypertext Documents. *Journal of the American Society fir Information Science, 47/12,* 896-908.
[22] Hackos, J.T. & D.M. Stevens (1997). *Standards for Online Communication: Publishing Information for The Internet/World Wide Web/Help Systems/Corporate Intranets*. New York: John Wiley & Sons.

---

1. For a discussion of evaluation functions I would like to refer to De Jong & Schellens [13].

2. I would like to thank J. Vanderhallen (UFSIA, Antwerp - Belgium), who helped me

with this study both practically and technically.

3. I would like to thank R. van Bodegom, E. de Boevère & E. Milder (KUB, Tilburg - The Netherlands), who collected the data for this study.

4. E. Faassen, I. Janssen & A. Vallen (KUB, Tilburg - The Netherlands, 1999, not published) investigated the characteristics of route descriptions on the Internet. They focused their analyses on style, structure, amount of information, interactivity and graphic design. The model they developed in their research paper was used to select the two sites for the second study.