APPLIED RESEARCH

SUMMARY

- Questions the contribution of heuristics to problem detection
- Shows that heuristics provide more focus to expert evaluation

Heuristic Web Site Evaluation: Exploring the Effects of Guidelines on Experts' Detection of Usability Problems

MARIEKE WELLE DONKER-KUIJER, MENNO DE JONG, AND LEO LENTZ

INTRODUCTION

When it comes to evaluating Web sites and other communication means, qualitative research strategies are the dominant approach. Rather than assessing the overall usability of a Web site or document, usability evaluators want to know in detail where usability problems occur and how these problems may be solved. Qualitative, troubleshooting methods help uncover these problems, either by carefully examining the usage process or by urging participants to make detailed and very specific comments (De Jong and Schellens 1997). A distinction can be made between expert-focused and user-focused methods (Schriver 1989). The majority of the methodological research thus far has focused on the validity of user-focused evaluation methods and on the comparison of expert-focused and user-focused approaches (cf. De Jong and Schellens 2000). Within the domain of expert-focused evaluation, various approaches have been developed to enhance the quality of expert evaluation, one of which is the use of heuristics. This article describes an empirical study combining quantitative and qualitative analyses to explore the pros and cons of heuristics as a qualitative approach of Web site evaluation.

Heuristic evaluation has become a popular method among Web site designers and usability professionals for assessing the quality of Web sites (Vredenburg and colleagues 2002). In a heuristic evaluation, experts systematically review a Web site by judging its compliance with recognized usability principles or guidelines (the heuristics) (Nielsen 1994). Originally developed for the evaluation of software applications, it is nowadays applied to all kinds of IT applications ranging from Web sites to virtual reality applications. Various sets of heuristics have been developed for many aspects of Web site quality, such as accessibility, usability, navigation, and comprehensibility (cf. De Jong and Van der Geest 2000). It is generally assumed that heuristics facilitate the evaluation process and enhance experts' skills in detecting usability problems in a Web site.

The evaluation process may be facilitated in various ways. Specific guidelines may complement the experts' own knowledge about the design of effective Web sites or may serve as mnemonic devices. Furthermore, the complete set of heuristics used may raise experts' awareness to primarily focus on the needs of potential users or to evaluate particular aspects of a Web site. For instance, experts working with heuristics about visual design may be expected to be more sensitive to visual presentation issues and see more problems in this area than experts in an unguided evaluation process. Heuristics may also support experts to systematically evaluate a Web site by offering a structured framework. A final advantage of heuristics is that they may be helpful in experts' communication about the evaluation results (Van der Geest and Spyridakis 2000). Nevertheless, little is known about the actual contribution of heuristics in the process of expert evaluation and about

Manuscript received 31 March 2008; revised 16 June 2008; accepted 18 June 2008.

the way experts incorporate the use of heuristics in their evaluation process. Despite the potential advantages mentioned above, heuristics also provide experts with the difficulty of having to switch between the heuristics and the Web site, thus complicating an already complex task.

Empirical research into the contribution of heuristics to the quality of expert evaluation has been limited. Studies by Sutcliffe (2002) and Paddison and Englefield (2004) found that experts are not always satisfied with the results of their heuristic evaluations. Sutcliffe also found that experts judged some of the heuristic items to be ambiguous when they had to be used in an actual evaluation. A recent study by Tao (2008) showed that information system professionals recognize and know many of the Web design guidelines available, but perceive difficulties in applying them to a specific Web site. Experts in a study by Hvannberg and colleagues (2007) mentioned the heuristic guidelines they had to work with both as a facilitator (seven times) and as a hindrance (six times) for detecting usability problems.

Only two studies explicitly addressed process characteristics of heuristic evaluation. Faulkner (2006) observed that experts do not necessarily use heuristics to identify usability problems in a Web site; some experts in her study primarily used the heuristics to label problems they had found solely relying on their own expertise. This might be indicative for the task complexity of evaluating a Web site when switching between heuristics and Web site is involved. The contribution of the heuristics to the detection of usability problems would be limited, and the requirement to label all problems afterward would possibly even have a negative effect on the number of usability problems identified. Apart from that, and on a more detailed level, Faulkner (2006) showed that the procedure of a heuristic evaluation may affect its effectiveness (in terms of the number of usability problems found). One half of the participants in her study worked for 40 min without breaks, whereas the other half worked in four 10-min blocks divided by 5-min breaks. Although some participants found the breaks distracting, the participants with breaks proved to be more productive in identifying usability problems. However, regardless of the set of heuristics used or whether or not the participants took breaks, work experience seemed to be the most important predictor of the number of problems detected, as was also found in studies by Nielsen (1992) and Saroyan (1993). A study by Hvannberg and colleagues (2007) focused on the effects of the medium of reporting problems (paper and pencil versus a Web-based registration tool) on the detection of usability problems and did not find significant differences between the two alternatives.

Several studies compared the results of heuristic evaluation with those of user-focused evaluation approaches, such as think-aloud usability testing (Bailey and colleagues 1992; Desurvire 1994; Fu and colleagues 2002; Hvannberg and colleagues 2007; Jeffries and colleagues 1991). These studies typically focused on the overlap in problems detected as well as on the total number and types of problems found. The results of these studies are mixed: in some cases, heuristic evaluation proved to be an effective way of detecting user problems; in other cases, heuristics only enabled experts to predict small parts of the (severe) usability problems.

Two studies compared evaluation results of experts working with and without heuristics, again with mixed results. Bastien and colleagues (1999) found that experts using one set of heuristics (the so-called Ergonomic Criteria) performed better than experts using another set of heuristics (the ISO/DIS 9241–10 dialog principles) and experts without any guidelines. Apparently, the effects of heuristics depend on the specific list of heuristic guidelines used. In the other study, by Connell and Hammond (1999), no differences were found between heuristic and unguided evaluations.

Surprisingly, no in-depth comparisons were made between unguided expert evaluations and heuristic evaluations. Furthermore, the empirical literature seems to neglect the potentially important distinction between high-level and low-level heuristics (cf. De Jong and Van der Geest 2000; Wright 1985). This distinction is comparable to the distinction between goal and action rules in the field of safety science (Hale and Swuste 1998). In the case of *high-level heuristics*, experts are given a limited number of more or less general guidelines, which are formulated as design aims rather than as specific design specifications. The guidelines define the goal to be achieved, without specifying how it should be achieved. An example is the advice to "work to ensure that users will view and notice links" (Farkas and Farkas 2000). High-level guidelines strongly rely on experts' professional knowledge to assess which design options are most suitable to achieve these aims. In the case of *low-level heuristics*, experts are given a larger set of detailed guidelines, which are formulated as design specifications rather than as design aims. The guidelines define a concrete action or a required state of the Web site. Low-level guidelines rely less strongly on experts' professional judgments about suitable design options, but instead are more likely to prescribe the desired action. "Well-established cues such as underlining and the raised 'button' appearance should be used to indicate links. Do not use these cues for other purposes," is an example of a low-level guideline (Farkas and Farkas 2000).

In this article, we present the results of a quantitative and qualitative study in which the contribution of heuristics was examined by a detailed comparison of experts' evaluation results with and without heuristics. The study involves a within-subjects design in which 16 participants first conducted an unguided evaluation, and after that, used a set of heuristics to evaluate a municipal Web site. One half of the participants worked with high-level heuristics and the other half with low-level heuristics. Our analysis focuses on the annotations (the problem detections and positive remarks) made by the experts. Specifically, the following questions are addressed:

- What does the unguided expert evaluation tell us about the validity of the heuristics?
- Are there any differences between heuristic and unguided expert evaluation regarding the number and types of annotations made?
- Do high-level and low-level heuristics have different effects on the annotations made by experts?

Our study contributes to the knowledge about heuristic evaluation in several respects. The results from the unguided evaluation are used to check whether the heuristics concerned actually cover all relevant aspects of navigation and comprehensibility. If experts implicitly use criteria in their unguided evaluation that are covered by the criteria in the heuristics, the heuristics can be said to reflect the knowledge in the field. A comparison of the numbers and types of annotations made in the unguided and the heuristic evaluation will shed light on the added value of heuristics for expert evaluations

The comparison of the effects of high-level and low-level heuristics is a first attempt to check whether this potentially important design feature of heuristics actually affects their usefulness. On the one hand, high-level heuristics may be more easily incorporated in the evaluation process (because they are easier to memorize and because it is easier to gain an overview of the complete set of heuristics) and may facilitate the detection of a broader range of usability problems (because of their goal instead of action orientation). On the other hand, low-level heuristics provide more specific cues for detecting usability problems.

MATERIALS AND METHODS

Participants

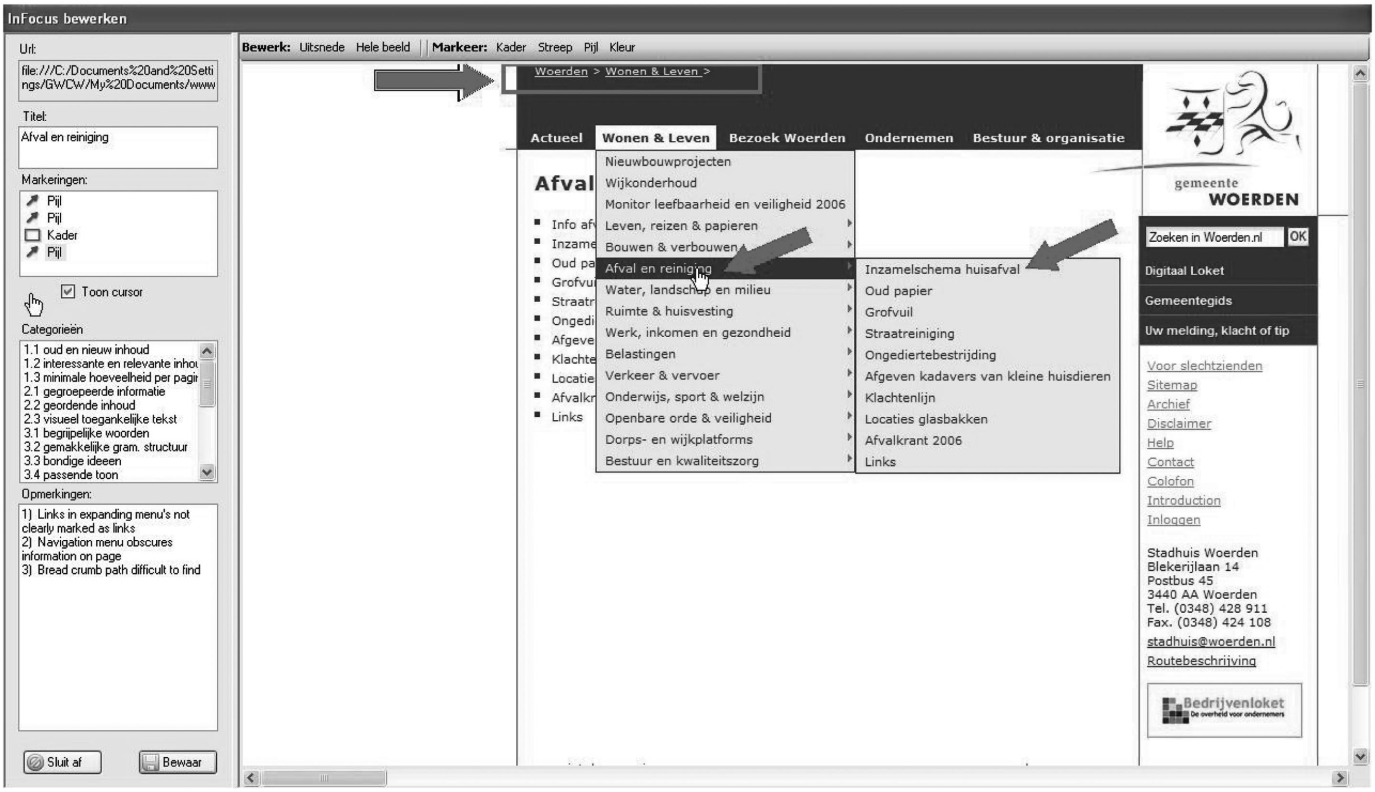
In this study, 16 communication professionals participated. Considering Saroyan's (1993) finding that differences in experts' background and perspective may lead to process and outcome differences, we kept the background of the participating experts as similar as possible. Experts were defined as communication professionals with a Master's degree in Communication Studies at the University of Twente and with at least 1 year of professional experience with designing and/or maintaining Web sites. Participants were approached through the university's alumni network. As an incentive, they received a gift voucher and a summary of the results of the study. Participants' professional experience ranged from 1 to 7 years (mean, 3.5 years). Most communication professionals worked as communication officers in commercial and noncommercial organizations (including municipalities), with responsibility for one or more Web sites. Four communication professionals were responsible for Web sites as part of another function (for example, a consultant in a small firm taking on the responsibility for a company Web site). Two communication professionals ran a commercial usability laboratory. The participants' age ranged from 25 to 36 years (mean, 28.5 years). Seven men and nine women participated.

Procedure

The communication professionals evaluated two parts of a municipal Web site. During the evaluation, they had to think aloud and record their positive and negative comments using Infocus, a software program for evaluating Web sites developed by Utrecht University. Infocus works as a normal Web browser but also offers the experts the opportunity to make screen shots; annotate them with boxes, lines, and arrows; and add explanatory text if they notice a problem or positive feature in the Web site (Figure 1). It is also possible for the research coordinator to offer evaluators a list of annotation categories to facilitate later analysis of the annotations.

First, the communication professionals performed an unguided evaluation for 25 min to assess their normal evaluation style. This evaluation was unguided in the sense that the communication professionals did not receive any (external) guidance as to how to perform the evaluation. The only instruction they received was to pay special attention to the navigation and comprehensibility of the Web site. In this session, they had to assign every annotation to the single category "General." After the unguided evaluation, they were presented with either high-level or low-level heuristics on the navigation and comprehensibility of Web sites (see Figure 2 and the subsection Heuristics for examples). They were first asked to read the heuristics and give a first impression about their usefulness. After that, the communication professionals evaluated a second part of the Web site using the heuristics (again for 25 min). In this session, they had to assign a heuristic category to each annotation. This evaluation was followed by a structured interview on their experiences and the annotations they had produced. In this article, we used the interviews mainly to solve unclear annotations. In a separate article, we will use the observation, think aloud, and interview data to further analyze the process characteristics of experts using heuristics.

The entire session lasted between 2 and 3 h. All communication professionals except two were asked to stop when they were still busy evaluating. Two of the communication professionals finished evaluating a section within the time: one during the unguided evaluation (18 annota-



InFocus bewerken

Url:
file:///C:/Documents%20and%20Settings/GWCW/My%20Documents/www

Titel:
Afval en reiniging

Markeringen:
- Pijl
- Pijl
- Kader
- Pijl

☑ Toon cursor

Categorieën
1.1 oud en nieuw inhoud
1.2 interessante en relevante inhou
1.3 minimale hoeveelheid per pagi
2.1 gegroepeerde informatie
2.2 geordende inhoud
2.3 visueel toegankelijke tekst
3.1 begrijpelijke woorden
3.2 gemakkelijke gram. structuur
3.3 bondige ideeen
3.4 passende toon

Opmerkingen:
1) Links in expanding menu's not clearly marked as links
2) Navigation menu obscures information on page
3) Bread crumb path difficult to find

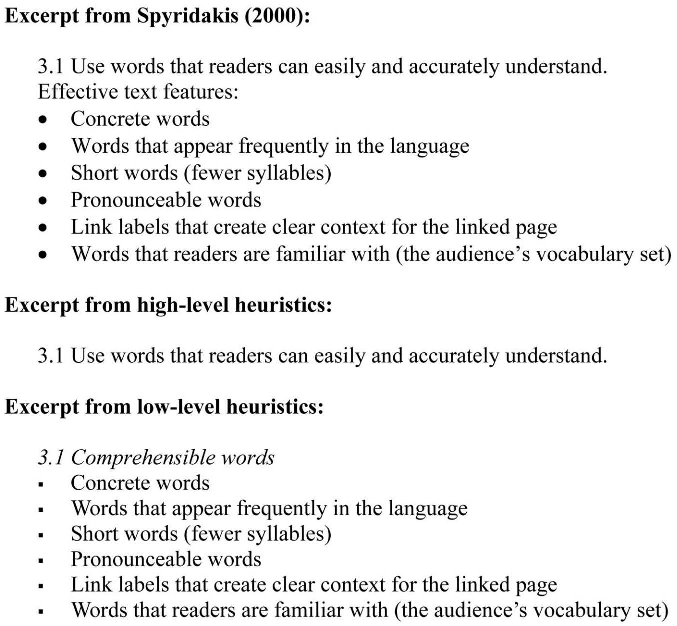
Figure 1. Example of an annotation screen in Infocus. Note: On the left side of the screen, one sees boxes that display the URL and title (="Titel") of the page, the markers used (="Markeringen"), the coding categories for the annotations, which in this case are from the items in the heuristics (="Categorieën"), and the text comments about the Web page (="Opmerkingen"). On the bottom left, the buttons for cancel (="Sluit af") and save (="Bewaar") are placed. The top bar contains buttons for different markers (="Markeer"), such as box (="Kader"), line (="Streep"), arrow (="Pijl"), and the color of these markers (="Kleur"). Note the arrows toward the unmarked links in the expanding navigation menu and the box around the bread crumb path.

tions) and the other during the heuristic evaluation (8 annotations). The sessions took place in quiet offices or (conference) rooms at the workplace, at home, or at the university.

Web site

The Web site used in this study was the municipal Web site of Woerden. The target group for this Web site is very broad, consisting of the citizens and resident organizations of the municipality and outside visitors interested in visiting the municipality for personal or professional reasons. The Web site contains information about all aspects of life in the municipality, ranging from policy information to passport applications and from municipal taxes to sightseeing spots. The Dutch government annually evaluates all Web sites by municipalities, provinces, boards of public works, and departments to see whether they conform to criteria of transparency, service quality, participation, and accessibility. The Web site of Woerden ranked 215th in the 2006 ranking for municipalities (total 485 places), which makes it an average quality municipal Web site. To ensure that the Web site remained the same during the research period, an offline copy was used.

Both sections used were part of the "Dwelling and Living" domain of the Web site. One section focused on "Garbage and Cleaning" and the other on "Education, Sports, and Well-being." These two sections were chosen because of their relative length—needed for the time duration of the evaluation—and their similar number of subpages. In addition, the topics were general enough for the experts not to need specialized expertise about the content. The order of the Web site parts alternated: one half of



Excerpt from Spyridakis (2000):

3.1 Use words that readers can easily and accurately understand.
Effective text features:

- Concrete words
- Words that appear frequently in the language
- Short words (fewer syllables)
- Pronounceable words
- Link labels that create clear context for the linked page
- Words that readers are familiar with (the audience's vocabulary set)

Excerpt from high-level heuristics:

3.1 Use words that readers can easily and accurately understand.

Excerpt from low-level heuristics:

3.1 Comprehensible words

- Concrete words
- Words that appear frequently in the language
- Short words (fewer syllables)
- Pronounceable words
- Link labels that create clear context for the linked page
- Words that readers are familiar with (the audience's vocabulary set)

Figure 2. Example of guidance regarding comprehensibility from Spyridakis (2000) and the corresponding comprehension item in the high-level and low-level heuristics.

the communication professionals conducted their unguided evaluation using the "Garbage and Cleaning" section, and the other half started with "Education, Sports, and Well-being."

Heuristics

The heuristics used were developed by combining two sets of research-based guidelines published in the autumn 2000 special issue of *Technical communication.* The first set offers advice on designing Web site navigation (Farkas and Farkas 2000) and the second focuses on the comprehensibility of Web pages (Spyridakis 2000). Of the comprehensibility heuristics, only the first three sections were used; to maintain a strong focus on comprehensibility, the sections on credibility and globalization were eliminated. Both sets of heuristics consist of general criteria formulated in one or two sentences, supplemented with several "key points" or "effective text features." The sentences can be seen as high-level heuristics, whereas the key points or effective text features are examples of low-level heuristics. To create a high-level set of heuristics, only the high-level sentences were used. In the low-level set of heuristics, the sentences were summarized in a short headline, and all the key points were presented underneath. The high-level heuristics consisted of one page. The low-level heuristics consisted of four pages. An example of high-level and low-level heuristics is given in Figure 2. This example illustrates why the length of the two heuristics differed so much. In low-level heuristics, many more items are needed to cover the same content as one item in high-level heuristics.

Analysis

The evaluation sessions were recorded and analyzed using the usability software program Morae and a webcam. Morae records the desktop activity, audio, and webcam video and synchronizes all into a single file, which can be further analyzed. We marked every change of page to assess which pages the communication professionals had seen, as well as the start of all Infocus annotations to assess when communication professionals decided to make an annotation. In addition, we used the Infocus output to analyze the content and nature of the annotations.

Although the communication professionals were asked to write down only one comment per annotation, some annotations did contain two or more comments. Therefore, all annotations were first checked for the occurrence of multiple comments. Two annotations were set aside, because they did not contain any comment on the Web site. This resulted in a set of 466 comments that formed the dataset for further analysis.

The first step in our data analysis involved the coding of the annotations made by the communication professionals. All comments were independently categorized by the first two authors. The coding scheme (Table 1) was initially based on the heuristics that the communication professionals had used. During the coding process, this scheme was deemed too limited, and three extra categories were added. These categories were "Comprehension general," "Navigation general," and "Other." The categories "Comprehension general" and "Navigation general" were added because we encountered annotations that clearly had to do with comprehension or navigation issues but were not covered by items in the heuristics. "Image supports text. Image gives more variation on the screen" is an example of such an annotation. The category "Other" was added to account for possible annotations that did not fit in with the navigation and comprehension heuristics at all—for example, "It may be a good idea to also offer the possibility to e-mail about complaints." All comments were also coded as either positive or negative. We used Cohen's kappa to assess the consistency of the work performed by the two coders. Cohen's kappa was 0.69 for the heuristic coding and 0.91 for the positive/negative coding, which indicates a satisfactory to almost perfect intercoder reliability (Landis and Koch 1977).

We ran a few statistical tests on the data to test whether differences in evaluator behavior were statistically significant.

TABLE 1: OVERVIEW OF HEURISTIC CODING SCHEME FOR ANNOTATIONS.

Level 1	Level 2	Level 3
Comprehension (C total)	Comprehension general (C general)	
	Selection and presentation of information (C1)	Presentation that facilitates orientation (C1.1)
		Selection of relevant and interesting information (C1.2)
		Limited amounts of information per page (C1.3)
	Organization of content on page (C2)	Grouping of content (C2.1)
		Logical order of content (C2.2)
		Visually accessible and scannable content (C2.3)
	Style and language (C3)	Comprehensible words (C3.1)
		Comprehensible syntax (C3.2)
		Conciseness (C3.3)
		Appropriate tone of voice (C3.4)
Navigation (N total)	Navigation general (N general)	
	Design of effective links (N1)	Links that are recognizable as links (N1.1)
		Noticeable links (N1.2)
		Clear link destinations (N1.3)
	Management of large numbers of links (N2)	Effective breadth/depth ratio in hierarchy (N2.1)
		Combination of primary and secondary links (N2.2)
		Appropriate converging of hierarchy branches (N2.3)
		Interface design that reveals underlying structure (N2.4)
	Provision of orientation information (N3)	Orientation information on home page (N3.1)
		Orientation information on lower pages (N3.2)
	Augmentation of link to link navigation (N4)	Sitemaps (N4.1)
		Search facilities (N4.2)
		Link to homepage on each page (N4.3)
Other (O total)		

◆ *The categories in level 3 were directly based on the heuristics. The categories Comprehension general, Navigation general and Other were added based on a first analysis of the annotations.*

We wanted to know, for example, whether the types of annotations produced during unguided evaluations are significantly different from the types of annotations produced during heuristic-based evaluations, and we wanted to know whether the annotations produced during evaluations using high-level heuristics are significantly different from the types of annotations produced during evaluations using low-level heuristics. Having examined similarities and differences between these different evaluation approaches, we moved to the heart of our analysis, which is the in-depth analysis of the content of annotations produced by the participating communication professionals. The statistical tests allowed us to consider whether different approaches are significantly different; the qualitative analysis allowed us to consider the nature of the differences and to delve into the specific results produced in the three different evaluation approaches that we considered.

RESULTS

We will first discuss the validity of the heuristics and then go into the comparison between unguided and heuristic evaluation. After that, we will address the comparison between high-level and low-level heuristics and describe in more detail the comments in both conditions on one particular Web page. Differences were tested using nonparametric tests. These statistical tests are most appropriate when the data come from a small sample and are not normally distributed, as was the case in this study.

Validity of the heuristics

An important criterion for the validity of heuristics is that they reflect state-of-the-art knowledge about effective Web site design (De Jong and Van der Geest 2000). The unguided evaluation data from our study can be used to check whether the heuristics indeed cover all (or most) relevant aspects of navigation and comprehensibility. To assess the heuristics' validity, we examined which percentage of all navigation and comprehension comments made by the communication professionals were covered by the specific guidelines in the two heuristics.

Of all comments regarding navigation, 69% had a clear relation with the guidelines offered in the heuristics. This means that the content of the annotations was very similar to the content, and sometimes even the wording, of an element in the heuristics. An example of this was the annotation "It is not clear to me what I can find behind this link. Is there more information? Or something else?," which is very similar to the N1.3 item of the navigation heuristics ("Be sure that all links clearly indicate their destinations").

Of all comments regarding comprehension, 86% corresponded to the guidelines in the heuristics. An example here is the annotation "Try to make the text a bit more scannable by separating the list from the other text," which even uses some of the wording of the C2.3 item of the comprehension heuristics ["Use organizational cues to make text visually accessible and scannable (easily skimmed or quickly read through at a top level); and to facilitate search tasks, comprehension, and recall. Do not distract readers with unnecessary cues"].

Combined, the heuristics covered 78% of all comprehension and navigation annotations made by the communication professionals. These figures confirm that the content of the heuristics matches the expert knowledge of communication professionals about navigation and comprehension and thereby confirm the relevance of these heuristics for experts who want to evaluate Web sites on these issues. At the same time, however, they question the novelty value of the heuristics: given the fact that communication professionals more or less naturally adopted many of the same evaluation criteria as those comprised in the heuristics, the heuristics' actual contribution to the detection of user problems cannot be expected to be very strong. If experts are already inclined to look at the visibility of links, getting the advice from the heuristics to look at this issue will not lead to additional problem detections.

An important omission in the navigation heuristics involved possible disorientation after users had clicked a certain link. The heuristics cover the clarity of the destination of links but do not sufficiently cover the requirements of the destination page. For instance, communication professionals were concerned about users' disorientation when a link opened a new screen, which seemed to replace the original Web site. They also criticized the fact that users landed smack dab in the middle of the destination place and were forced to start looking for the desired information all over again. An omission in the comprehension heuristics involved the possible contribution of images to help the users visualize and understand the information offered. One example is: "Now that I've seen page X with a photo, it might be handy to add a picture of the blue garbage container on this page, so that everybody knows what is meant by that." This and similar annotations might warrant an extra subcategory in the heuristic, "Visual support for information."

Comparison between unguided and heuristic evaluation

In total, the communication professionals wrote down 466 annotations: 269 in the unguided evaluation and 197 in the heuristic evaluation. The communication professionals differed in the number of comments they made within the same time frame. The mean number of comments in the unguided evaluation was 16.8 (range, 6 to 31), and the mean number of problems in the heuristic evaluation was 12.3 (range, 6 to 28). The trend is toward a lower number of comments in the heuristic evaluation (Wilcoxon signed ranks test, $Z = -2.93$, $P < 0.005$; this test result indicates that it is unlikely that the trend is accidental or random but rather that it is statistically significant). The Morae footage shows that this difference may be attributed to the time communication professionals needed to find the appropriate heuristics to categorize their comments. This categorizing time seems to have slowed them down. In general, this may mean that experts performing a heuristic evaluation for the first time need more time to complete it.

Subdivision in positive and negative annotations between evaluation conditions

Heuristic evaluation may yield both positive and negative comments on a Web site, because it also directs evaluators' attention to areas without problems ("There are no problems with language use" translates into "The language use is good."). In this study, it is interesting to see that only those communication professionals that created positive annotations in the unguided evaluation also created positive annotations in the second round. The number of negative annotations was significantly lower in the heuristic evaluation (10.3) than in the unguided evaluation (13.9; Wilcoxon signed ranks test, $Z = -3.22$, $P < 0.01$). There were no differences in the number of positive annotations between the unguided and heuristic evaluation (Wilcoxon signed ranks test, $Z = -1.01$, not significant).

Subdivision in heuristic categories between evaluation conditions

If we look at the level 1 coding—that is, the subdivision over the three categories Comprehension, Navigation, and Other—significant differences were found between the unguided and heuristic evaluation in the number of annotations within the categories Navigation and Other. The number of "Other" annotations fell from 3.3 in the unguided condition to 1.4 in the heuristic condition (Wilcoxon signed ranks test, $Z = -2.05$, $P < 0.05$). Surprisingly, however, the mean number of "Navigation" annotations also dropped from 5.5 to 3.7 in the heuristic evaluation (Wilcoxon signed ranks test, $Z = -2.11$, $P < 0.05$). The number of annotations in the category "Comprehension" did not differ between the two evaluation conditions (Wilcoxon signed ranks test, $Z = -0.83$, not significant).

On the second level, there were no differences in the number of "Comprehension" annotations per criterion between the two evaluation conditions. Within the "Navigation" annotations, however, one significant difference was found: the number of navigation problems that did not correspond to the specific navigation criteria included in the heuristics dropped from 1.8 in the unguided evaluation to 0.7 in the heuristic evaluation (Wilcoxon signed ranks test, $Z = -2.48$, $P < 0.05$). For the other second level "Navigation" annotations, no differences between the evaluation modes were found.

The number of annotations in all the third level categories did not differ between the unguided and heuristic evaluations, except for one category. These annotations involved the heuristic C1.2 ("Selection of relevant and interesting information"). In the heuristic evaluation, the communication professionals created significantly less annotations in this category than in the unguided evaluation: 2.8 in the unguided evaluation versus 1.5 in the heuristic evaluation (Wilcoxon signed ranks test, $Z = -1.99$, $P < 0.05$).

If we look at the content of these C1.2 annotations, they are usually very specific, asking for more information regarding the topic of the page. An example is: "Now that I have found the information [about the city dump], I miss information about whether the garbage needs to be presented/packaged in a special way or not." In the heuristic evaluation, we find several more general annotations that copy the wording of this heuristic but that do not go into details. Examples are "I miss relevant information" and "Residents find this interesting and useful information." In the second case, the lack of specificity is not a problem, but a designer needing to remedy the first annotation will have problems deciding what needs to be added.

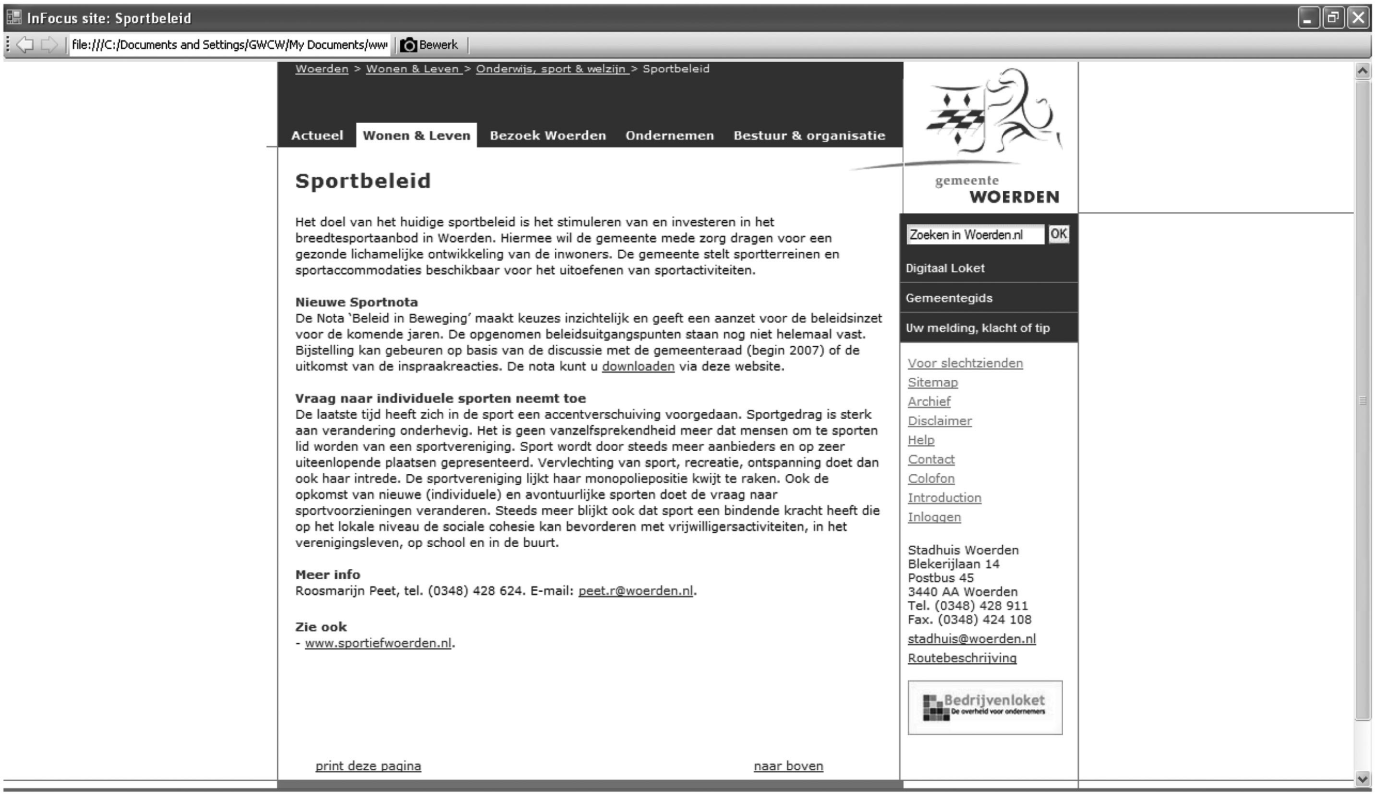
Comparison of high-level and low-level heuristics

No differences were found between the high-level and low-level heuristics in the number of annotations, the subdivision in positive and negative annotations, and the subdivision in the different heuristic categories.

In-depth analysis of annotations

The general trend seen in the previous section was that, under the influence of the heuristics, the annotations were increasingly focused on navigation and comprehension issues covered by these heuristics. To see what this effect does in actual annotations, we will now take a closer look at the annotations regarding one page all communication professionals evaluated: the "Sports Policy" Web page. This page is a part of the "Education, Sports, and Well-being" section and is listed at the top of the overview page and the navigation menu, so all communication professionals were likely to see this page early in their evaluations. It contains a medium amount of content and has some secondary links to related information both in the text and at the bottom. General secondary links, such as contact, sitemap, accessibility, and the link to the English language site are listed on the right side of the page (Figure 3). Most of the heuristics could be applied to this page.

In all, the communication professionals made 29 annotations in the unguided evaluation and 39 annotations in the heuristic evaluation. Similar to the general trend of increasing specificity, the number of "Other" annotations dropped from 9 to 2 and the number of general "Comprehension" (C-general) and "Navigation" (N-general) annota-



Sportbeleid

Het doel van het huidige sportbeleid is het stimuleren van en investeren in het breedtesportaanbod in Woerden. Hiermee wil de gemeente mede zorg dragen voor een gezonde lichamelijke ontwikkeling van de inwoners. De gemeente stelt sportterreinen en sportaccommodaties beschikbaar voor het uitoefenen van sportactiviteiten.

Nieuwe Sportnota
De Nota 'Beleid in Beweging' maakt keuzes inzichtelijk en geeft een aanzet voor de beleidsinzet voor de komende jaren. De opgenomen beleidsuitgangspunten staan nog niet helemaal vast. Bijstelling kan gebeuren op basis van de discussie met de gemeenteraad (begin 2007) of de uitkomst van de inspraakreacties. De nota kunt u downloaden via deze website.

Vraag naar individuele sporten neemt toe
De laatste tijd heeft zich in de sport een accentverschuiving voorgedaan. Sportgedrag is sterk aan verandering onderhevig. Het is geen vanzelfsprekendheid meer dat mensen om te sporten lid worden van een sportvereniging. Sport wordt door steeds meer aanbieders en op zeer uiteenlopende plaatsen gepresenteerd. Vervlechting van sport, recreatie, ontspanning doet dan ook haar intrede. De sportvereniging lijkt haar monopoliepositie kwijt te raken. Ook de opkomst van nieuwe (individuele) en avontuurlijke sporten doet de vraag naar sportvoorzieningen veranderen. Steeds meer blijkt ook dat sport een bindende kracht heeft die op het lokale niveau de sociale cohesie kan bevorderen met vrijwilligersactiviteiten, in het verenigingsleven, op school en in de buurt.

Meer info
Roosmarijn Peet, tel. (0348) 428 624. E-mail: peet.r@woerden.nl.

Zie ook
- www.sportiefwoerden.nl.

Stadhuis Woerden
Blekerijlaan 14
Postbus 45
3440 AA Woerden
Tel. (0348) 428 911
Fax. (0348) 424 108
stadhuis@woerden.nl
Routebeschrijving

Figure 3. Screen shot of "Sports policy" Web page in Infocus browser. Note: The arrows in the top bar are backward and forward buttons that function just like similar buttons in other browsers. The central white box contains the URL of the current page and can be used to navigate to other pages by typing in an address. The camera/annotation button (="Bewerk") is placed on the right. Clicking on this button opens the annotation screen as seen in Figure 1. The camera is symbolic for the screen shot Infocus makes of the current Web page and that can subsequently be annotated in the special annotation screen.

tions dropped from 3 to 0. In contrast with this decrease in general annotations, the number of annotations that corresponded directly to specific heuristics rose considerably. The number of annotations that had to do with the selection and presentation of content (C1) rose from 6 to 10 and the annotations about style and language (C3) increased from 1 to 10. There was also an increase in the number of navigation annotations that referred to the design of effective links (N1), which went from 3 to 7.

The drop in the number of "Other" annotations between the two conditions was consistent with the general picture of this study, but the content of these annotations also became more specific. The subjects of the "Other" annotations in the unguided evaluation showed a wider variation. They often regarded the general secondary links on the right side of the screen ("The text button for partially sighted people is too small") and aspects of the layout ("Clear, readable, no strange colors in text or background" and "The text itself is also relatively small"). In contrast, the only two "Other" annotations in the heuristic evaluation were "Well-organized site, not crowded, no banners, etc." and "It would be convenient to open this type of downloads in a new window, so that when you close the download, you do not close the site of woerden.nl." The content of these two annotations is much closer to the subject of the heuristics. The heuristics mention that links on a Web page should be noticeable and that the text should be visually accessible and scannable. Both aims can be achieved with well-structured pages that are not overcrowded. Regarding the second annotation, the heuristics pay attention to issues such as clear indications of the destination of external and internal links. Saying that downloads should open in a different window is not a big leap from that guideline.

In the content of the "Comprehension" annotations, the influence of the heuristics can also be seen. Six annotations in the unguided evaluation dealt with the selection and presentation of the content (C1). Comments were made about the amount of text on the page (three annotations), its relevance to the user (two annotations), and the desirability of publishing the address of and directions to the town hall on every page of the site (one annotation). Interestingly, one communication professional thought that there was "quite a lot of text to read," whereas two communication professionals said, about the same page, it was "short: on one page, I like that." The two communication professionals commenting on the relevance of the text, both wondered about "the added value" and "the relevance" of the information "for the citizen." In the heuristic evaluation, similar concerns about the length of the text (three annotations) and the relevance for the citizen (four annotations) were raised. As in the unguided evaluation, the communication professionals in the heuristic evaluation did not agree about the length of the text. One commented "Short page, not a long-winded story," whereas number 2 thought that "this is about the maximum amount of information I would put on a Web page," and number 3 said "This story looks like it is too long. I would shorten it or present a list in between." The communication professionals making annotations about the relevance showed a stronger agreement in their doubts regarding the relevance of the information for citizens of Woerden.

In addition to the annotations about length and relevance, three communication professionals were not satisfied with the content on the page and asked for more information about specific subjects ("What are the consequences? What are the changes?").

The picture of increased specificity in the annotations is even clearer when we look at the annotations that deal with style and language (C3). In the unguided evaluation, only one communication professional commented that there was "maybe a superfluous word" on the Web page. In the heuristic evaluation, however, the communication professionals saw many more problems with the style and language. There were problems with "difficult words" (four annotations) and "difficult syntax" (two annotations), the text contained "unnecessary details" (one annotation), and the communication professionals "would choose another tone to address inhabitants; it is now much too official" (three annotations). These examples show that the communication professionals suddenly seemed much more focused on the selection and presentation of information and the style and language used. The use of the same terms in both the annotations and the heuristics are another indication of this influence.

If we look at the annotations regarding the design of effective links (N1), two of the communication professionals in the unguided evaluation commented on the quality of the secondary links on the right hand side of the screen. Their annotations read:

- "Certainly the navigation at the right side raises some questions because of general terms like 'digital counter' and 'municipal guidebook' (does not refer to content, more to services)."
- "Introduction, is meant for English people. I don't think it's very logical. I would make it more eye-catching that this is the English version."

On the other hand, another communication professional stated that there was "structure in the information: links are clear." In the heuristic evaluation, the annotations were a bit more specific: five annotations regarded a link that did not state its destination. The complaints ranged from "It is not clear this link starts a dialog for downloading" to "Preferably indicate the size of the file that can be downloaded" and from "Is this a site from the municipality or a commercial site or the like; where will I end up or what does this site mention?" to "Use link labels." In contrast, judging by the annotations "The links are clear" and "Here, links are clearly accentuated," two communication professionals were positive about the quality of the links. On the basis of the annotations given, it seems the communication professionals were already aware of the necessity of recognizable links with a clear, unambiguous name but that the heuristics reinforced this awareness and alerted them to the possibility of adding more destination information than just an informative link name. Under the influence of the heuristics, they were more specific in their annotations. Also, as was the case with the comprehension annotations, the influence of the navigation heuristics was visible in the use of the same terminology, such as link labels and download dialog.

To summarize, this analysis of the exact wording of annotations has shown how the general trends from the more general analyses can also be seen in the content of very specific annotations. The content of the "Other" annotations had more bearing on the subject of the heuristics, and the formulation of the comprehension and navigation annotations was more specific, sometimes even using the same terminology as the heuristics.

DISCUSSION

The results of our study confirm the validity of the heuristics on navigation (Farkas and Farkas 2000) and comprehensibility (Spyridakis 2000): both heuristics strongly reflect the state-of-the-art knowledge that the communication professionals in our study brought to the unguided evaluation task. At the same time, however, our results raise questions about the practical usefulness of the heuristics in this particular setting. A remarkable result is the decrease in the number of annotations between the unguided and heuristic evaluation, because the communication professionals needed time to find the appropriate heuristic to label the annotation.

This does not necessarily correspond to a structural disadvantage of heuristics but underlines two unfavorable aspects of the heuristic evaluation under study, which must be considered by organizations and professionals who think of adopting heuristic evaluation. First, the practical usefulness of heuristics depends on their validity and novelty value, and there is often a tension between the two criteria. The heuristics used in this study proved to strongly reflect the knowledge in the field about navigation and comprehensibility but had relatively little novelty value for this particular group of communication professionals. The usefulness of heuristics can only be assessed by considering the prior knowledge and evaluation practices that communication professionals already have. Second, the practical usefulness of heuristics will probably enhance when communication professionals become more experienced with them. Heuristics will probably be more beneficial when organizations or experts have adopted them as a standard evaluation procedure for many Web sites and/or when they are introduced in an educational program.

Another important effect of the heuristics is that they seemed to focus experts' attention to navigation and comprehensibility. The heuristics seem to cause communication professionals to limit their attention to issues that fit within the framework of the heuristics. Two results are indicative for this phenomenon. One indication is a drop in the category "Other" annotations between the unguided and the heuristic evaluation. The use of the heuristics did not lead to the detection of more navigation and comprehension problems but to a decrease of problem detections that did not correspond to "Navigation" and "Comprehension." The heuristics narrowed the experts' attention for user problems, or, more positively formulated, gave them a clearer focus. Another, more unfavorable, indication is the decrease of navigation annotations that did not correspond to the specific navigation guidelines. Working with the heuristics narrowed the communication professionals' views on the various aspects of user-friendly navigation.

The distinction between high-level and low-level heuristics did not seem to have any impact on the annotations of the communication professionals in this study. We can only speculate on the reasons for this finding. This might be because of the short period between the introduction and the use of the heuristics. Maybe with a longer training period or a longer duration of the evaluation, the influence of the type of heuristics will be more pronounced.

Limitations of the study

This study has two limitations. First, the experts had to use the heuristics immediately after receiving them. They had difficulty finding a modus operandi in integrating the heuristics into their evaluation process. Maybe with more time, the use of heuristics would have become more natural for them, and they would have had the chance to internalize (parts of the) heuristics. This might speed up the evaluation process and lead to more, and more diverse, annotations. In future research, we will therefore focus on more structural ways of using heuristics to evaluate Web sites. Second, the assignment to think aloud may have affected the communication professionals' evaluation process in both conditions. It is not likely that the cognitive load on the professionals was too high, because the evaluation of Web sites was not a very novel or complex task for them. However, the assignment to think aloud probably slowed down the evaluation process and might have urged participants to work more systematically than they would have done in normal evaluation settings.

Practical implications

Practitioners who consider conducting a heuristic evaluation of a Web site for the first time need to be aware that this may take up more time and energy than an unguided evaluation from their own expertise. In addition, the subject of the heuristics needs to be chosen carefully, because their focus will be mostly limited to this subject. However, if professionals find a modus operandi for incorporating the heuristics in their work process, heuristic evaluation may be a valuable method. Such a modus operandi could be a combination of the unguided and heuristic evaluation, in which each page is first scanned for problems that "jump to the eye" and subsequently evaluated according to the (order of the) heuristics. In addition, after the evaluation is completed, the annotations of both unguided and heuristic evaluations can be categorized according to the heuristics to achieve a more structured discussion of the results. TC

ACKNOWLEDGEMENTS

This paper is from a research project that was financed by the Dutch Organization for Scientific Research. It is part of the research program Evaluation of Municipal Websites. The authors thank guest editor James Conklin and two anonymous reviewers for constructive and useful suggestions to improve this article.

REFERENCES

Bailey, R. W., R. W. Allan, and P. Raiello. 1992. Usability testing vs. heuristic evaluation: A head-to-head comparison. Proceedings of the Human Factors and Ergonomics Society, 36th Annual Meeting, pp. 409–413. Santa Monica, CA: HFES.

Bastien, J.M.C., D. L. Scapin, and C. Leulier. 1999. The ergonomic criteria and the ISO/DIS 9241–10 dialogue principles: A pilot comparison in an evaluation task. *Interacting with computers* 11:299–322.

Connell, I. W., and N. V. Hammond. 1999. Comparing usability evaluation principles with heuristics: Problem instances versus problem types. In *Human-computer interaction—INTERACT '99*, Ed. M. A. Sasse and C. Johnson.Amsterdam: IOS Press, pp. 621–636.

De Jong, M., and P. J. Schellens. 1997. Reader-focused text evaluation: An overview of goals and methods. *Journal of business and technical communication* 11:402–432.

——— and ———. 2000. Toward a document evaluation methodology. What does research tell us about the validity and reliability of evaluation methods? *IEEE transactions on professional communication* 43:242–260.

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

Desurvire, H. W. 1994. Faster, cheaper!! Are usability inspection methods as effective as empirical testing? In *Usability inspection methods,* Ed. J. Nielsen and R. L. Mack. New York: John Wiley, pp. 173–202.

Farkas, D. K., and J. B. Farkas. 2000. Guidelines for designing Web navigation. *Technical communication* 47:341–358.

Faulkner, L. L. 2006. *Structured software usability evaluation: An experiment in evaluation design.* PhD diss., University of Texas at Austin.

Fu, L., G. Salvendy, and L .Turley. 2002. Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & information technology* 21:137–143.

Hale, A. R., and P. Swuste. 1998. Safety rules: Procedural freedom or action constraint? *Safety science* 29:163–177.

Hvannberg, E. T., E.L.-C. Law, and M. K. Lárusdóttir. 2007. Heuristic evaluation: Comparing ways of finding and reporting usability problems. *Interacting with computers* 19:225–240.

Jeffries, R. M., J. R. Miller, C. Wharton, and K. Uyeda. 1991. User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of the SIGCHI conference on human factors in computing systems: Reaching through technology*, ed. S.P. Robertson, G.M. Olson, and J.S. Olson. New York: ACM, pp. 119–124.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33: 159–174.

Nielsen, J. 1992. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, ed. P. Bauersfeld, J. Bennet, and G. Lynch. New York: ACM, pp. 373–380.

———. 1994. Heuristic evaluation. In *Usability inspection methods,* Ed. J. Nielsen and R. L. Mack. New York: John Wiley, pp. 25–62.

Paddison, C., and P. Englefield. 2004. Applying heuristics to accessibility inspections. *Interacting with computers* 16: 507–521.

Saroyan, A. 1993. Differences in expert practice: A case from formative evaluation. *Instructional science* 21:451–472.

Schriver, K. A. 1989. Evaluating text quality: The continuum from text-focused to reader-focused methods. *IEEE transactions on professional communication* 32:238–255.

Spyridakis, J. H. 2000. Guidelines for authoring comprehensible Web pages and evaluating their success. *Technical communication* 47:359–382.

Sutcliffe, A. 2002. Assessing the reliability of heuristic evaluation for website attractiveness and usability. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02),* Vol. 5. Washington, DC: IEEE Computer Society, pp. 137–146.

Tao, Y.-H. 2008. Information system professionals' knowledge and application gaps toward Web design guidelines. *Computers in human behavior* 24:956–968.

Van der Geest, T., and J. H. Spyridakis. 2000. Developing heuristics for Web communication: An introduction to this special issue. *Technical communication* 47:301–310.

Vredenburg, K., J.-Y. Mao, P. W. Smith, and T. Carey. 2002. A survey of user-centered design practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* Minneapolis, MN: ACM, pp. 471–478.

Wright, P. 1985. Editing: Policies and processes. In *Designing usable texts,* Ed. T. M. Duffy and R. Waller. Orlando, FL: Academic Press, pp. 63–96.

MARIEKE WELLE DONKER-KUIJER is a PhD candidate at the University of Twente, Twente, The Netherlands. Her PhD research focuses on the methodology of evaluating informative Web sites, more specifically on the merits and drawbacks of heuristic expert evaluation. Contact: m.c.j.welledonker-kuijer@gw.utwente.nl.

MENNO DE JONG is an associate professor of communication studies at the University of Twente, Twente, The Netherlands. His main research interest concerns the methodology of applied communication research. Contact: m.d.t.dejong@utwente.nl.

LEO LENTZ is an associate professor of communication studies at Utrecht University, Utrecht, The Netherlands. Web site usability and document design are the main focuses of his research. Contact: l.lentz@let.uu.nl.