Website Evaluation Questionnaire: Development of a Research-Based Tool for Evaluating Informational Websites

Sanne Elling[1], Leo Lentz[1], and Menno de Jong[2]

[1] Utrecht University, Utrecht Institute of Linguistics (UIL-OTS), Trans 10,
3512 JK Utrecht, The Netherlands
s.k.elling@let.uu.nl, l.lentz@let.uu.nl
[2] University of Twente, Faculty of Behavioral Sciences, Institute for Behavioral Research,
P.O. Box 217, 7500 AE Enschede, The Netherlands
m.d.t.dejong@utwente.nl

Abstract. Online questionnaires are frequently used to monitor the quality of municipal and other governmental websites. In the present situation, many government organizations seem to reinvent the wheel and develop their own questionnaire. This leads to the undesirable situation that website quality is often assessed with instruments that are not comparable with each other and are not empirically validated. This article presents a generic Website Evaluation Questionnaire (WEQ) for the evaluation of informational websites. The WEQ was developed on the basis of the literature on usability and user satisfaction and was tested and revised in several rounds. This has resulted in a reliable questionnaire measuring clearly distinct quality dimensions of informational websites. The WEQ can be used by governmental organizations for evaluating their websites and for benchmarking their results against each other.

Keywords: Website design, website evaluation, questionnaire, website usability.

1 Introduction

The pressure on governmental bodies to develop websites that enable citizens to participate in a modern democracy has reached high proportions [1]. Governments do not only provide information to their residents but increasingly use their websites to facilitate interaction and offer online services to national and international audiences. Websites have evolved to important information and service channels between governmental organizations and citizens and other stakeholders. Evaluation research is necessary to monitor and further improve the quality of these websites. Several expert-focused and user-focused methods are available for this type of evaluation research, of which (heuristic) expert evaluation and think-aloud usability testing are the most current laboratory approaches. These approaches typically produce detailed and diagnostic feedback, which may be used to revise a website or certain web pages.

A more coarse-grained evaluation method which focuses predominantly on the overall quality of websites is the online questionnaire. Many governmental and other

M.A. Wimmer, H.J. Scholl, and A. Grönlund (Eds.): EGOV 2007, LNCS 4656, pp. 293–304, 2007.

organizations use such a questionnaire to collect feedback on their website from real visitors. Online questionnaires are a cheap and easy way of gathering user feedback. Most of these organizations develop their own evaluation questionnaire, which has the potential advantage that the questions asked may be tailored to the specific characteristics of the website, but also has two important drawbacks. First, the evaluation results of governmental websites cannot be compared to each other, due to differences between the questionnaires used. Second, the validity and reliability of all individually developed questionnaires is questionable or at best unknown.

In this paper, we will describe a project aimed at developing and validating a generic Website Evaluation Questionnaire (WEQ), which may be used to evaluate municipal and other governmental websites. We will first address the criteria for a methodologically sound questionnaire. After that, we will discuss previous, more general, web evaluation questionnaires available in the literature. Then we will outline the design of the WEQ and describe the various studies we conducted to assess and improve its validity and reliability. The WEQ itself can be found in the Appendix.

2 Validity and Reliability of Web Evaluation Questionnaires

On the internet, many examples of problematic questionnaires can be found, underlining that designing a good survey is not an easy task. It is all about identifying the relevant constructs to be measured and asking the right sets of questions to measure them. We will discuss three important topics concerning validity.

- Which definition of website quality is used?
- How do the results of a questionnaire relate to the respondents' experiences when using the website?
- How does the group of questionnaire respondents relate to the website's overall target audience?

The first important issue is the definition of website quality. There is no agreement about the question what website quality exactly is and which dimensions or items a questionnaire should contain. In the case of informative websites, it seems plausible to connect website quality to usability. Nielsen & Loracher [2] define the concept of usability as follows.

> '… a quality attribute relating to how easy something is to use. More specifically, it refers to how quickly people can learn to use something, how efficient they are while using it, how memorable it is, how error-prone it is, and how much users like using it.' (Nielsen & Loracher [2] p.xvi).

This is a rather broad focus which relates to a wide range of (specific) usability guidelines as presented in their recently published book, varying from the optimal place to put links, to choosing fonts, to tips for the right place to display prices. In this definition, three notions of the ISO standard can be found: effectiveness, efficiency and user satisfaction [3]. The definitions of Nielsen and ISO are most frequently referred to in the literature on website usability.

In a review of 180 studies on usability, Hornbaek distinguishes between subjective and objective measures of usability—e.g., perceptions of task difficulty (subjective) and usage patterns (objective) [4]. The aspects of website quality that can be measured using a questionnaire are limited to the subjective experiences of visitors. Visitors may have opinions about the website itself, about the process of using it, and about the outcomes of their interactions with the website. Their opinions about the process relate to the navigation process and the accessibility of information. The visitors' opinions about the outcome concern the quality of the information found.

The second issue is whether the task of filling out an evaluation questionnaire really reflects the opinions visitors had when using the website. The process of answering a questionnaire is complex and may lead to biases. Sudman et al. [5] give an overview of the tasks respondents must perform when answering questions. They must first interpret the question and understand its meaning. If the question involves an opinion, respondents must retrieve a previously formed opinion from memory or decide on an opinion at the very moment. To form an opinion, they need to make a mental representation of the artifact they are to evaluate and retrieve or construct a standard against which it can be evaluated. Then their opinion must be communicated to the researcher, often after formatting the response to fit to the response alternatives provided with the question. A common bias in usability research, which we also found in our pilot studies, is that people tend to be more positive in a questionnaire then would be justified considering the usability problems they have encountered. It is imaginable that people filling out a questionnaire have forgotten many of their problems using the website, and that the questionnaire creates new attitudes that respondents were not aware of during navigation.

The third issue is the representativeness of the sample of respondents. Couper [6] discusses two problems that are important for governmental website evaluation. The *sampling error* is the problem that not every member of the population has the same chance to be included in the survey. An example of this error is that people who enter the website via other routes than the homepage may not see the survey when it is only shown on the homepage. Another problem is the *nonresponse error*, which means that not everyone in the target group will be inclined to participate. For example, a lack of time, a negative attitude toward the organization or technical problems can keep people from filling out the questionnaire, which may lead to a non-representative sample. Little is known about ways of motivating people to take part in a web survey. Dillman & Bowker [7] present some advice for motivating people, but they point out that there is only little or no experimental evidence and underline the need for more research on this topic.

Having discussed three aspects of validity, we will finish this section with discussing the reliability of questionnaires. In the context of this paper we concentrate on the idea of item-reliability. This involves the question whether website quality dimensions are measured in a consistent way. Items that are supposed to measure the same dimension should have a Cronbach's alpha of at least .70. Low reliability scores can be caused by difficult or ambiguous formulations. Molenaar [8] gives an overview of several types of such formulations and their effects on the responses.

3 Previous Questionnaires on Website Quality

In the literature, we found three earlier research projects focusing on the systematic development and validation of website evaluation questionnaires. We analyzed these studies with the purpose of defining dimensions of website quality for the WEQ. The analysis focused on dimensions that relate to the navigation process and dimensions concerning the quality of the information.

Kirakowski [9] describes the Website Analysis Measurement Inventory (WAMMI), a questionnaire consisting of 60 questions, which have to be answered on seven-point Likert scales. The concept of website usability is divided into five categories. The degree to which users:

- feel efficient
- like the system
- find the system helpful
- feel in control of the interactions
- can learn to use the system

These five categories are the result of an analysis of the feedback that was produced by a large group of website designers and users. Kirakowski reports high Cronbach's alphas (between 0.70 and 0.90) for the dimensions. For practical use, the WAMMI questionnaire has been reduced to a set of 20 questions, which place less of a burden on the respondent. The first four dimensions are for the most part related to the users' attitude towards the website and the process of interaction. The last category of *learnability* presupposes that the site will be visited repeatedly by its audience. In the context of governmental websites, we think this category to be less relevant, since the low frequency of citizen visits will not allow them to really learn to use the site.

Van Schaik and Ling [10] developed another evaluation questionnaire, which also consisted of five categories. Their dimensions are:

- perceived ease of use
- disorientation
- flow
- perceived usefulness
- aesthetic quality

Respondents visited a university website and performed three information retrieval tasks. After that, they filled out the questionnaire, which consisted of 30 questions. The authors report high scores on the Cronbach's alpha (between 0.74 en 0.89). In a post-hoc analysis they decided to split the *flow* dimension into two sub dimensions: *involvement* and *control*. The first three categories are clearly related to attitudes towards the interaction process. The *perceived usefulness* seems to be related to attitudes towards the outcome of the process. A new category is the *aesthetic quality*, which focuses on the general appearance of the website itself.

In our view, the dimension of *flow* is less relevant in the context of governmental websites. *Flow* is defined as a psychological condition in which a person feels cognitively efficient, motivated and happy. Citizens that visit websites in order to find out how to get a new passport or to inform the local authority about a change in their address, will not expect these sites to create a feeling of flow.

According to Lavie and Tractinsky [11] the aesthetic dimension may be divided into a notion of classical aesthetics (a clear, clean, symmetric and pleasant design) and expressive aesthetics (creative, fascinating and original design). They found a clear correlation between the first notion and attitudes towards the usability of a website. This would mean that a "classically designed" website helps people to better perform their tasks. For governmental websites this notion might be relevant. We do not think visitors expect these sites to be original and fascinating, so the second notion of aesthetics will not be incorporated into the WEQ.

Muylle et al. [12] developed the WUS (Website User Satisfaction questionnaire). This 60 item questionnaire consists of four main dimensions of user satisfaction and eleven sub dimensions. A sample of 837 website users filled out this questionnaire after having visited a site of their own choice. The authors report high reliability rates (between 0.74 and 0.89). A confirmatory factor analysis supported the distinction in four main dimensions and eleven sub dimensions:

- connection
    - ease of use
    - entry guidance
    - structure
    - hyperlink connotation
    - speed
- quality of information
    - relevance
    - accuracy
    - comprehensibility
    - comprehensiveness
- layout
- language

The first dimension of *connection* clearly is related to the users' attitudes towards the interaction process. The second dimension q*uality of information* is related to outcome attitudes. The *layout* dimension is strongly connected to the aesthetic quality in the classical notion that we discussed above. The *language* dimension is defined as the degree to which the choice of the language of communication is tailored to the user. In multilingual countries like Belgium, this may be a relevant aspect. For the questionnaire we developed, it seems more useful to aim a *language* dimension at the comprehensibility of the language use on the website.

For the development of the WEQ we concentrated on three dimensions: the attitudes towards the interaction process, the attitudes towards the outcome of the process and the attitudes towards the classical aesthetics. Our starting point was the WUS, because this questionnaire focuses more than the other two on users who are searching for information on a website. Moreover, the WUS pays a lot of attention to the quality of information, which we consider highly relevant for the domain of municipal websites. There are two major changes between the WUS and the first version of our questionnaire. The first change concerns the *language* dimension. We transformed the questions about language choice into questions about the language use in the website and put this as a sub dimension under *quality of information*. The second change was the introduction of a new sub dimension in the *connection* section with questions about the search engine. We consider this to be an important tool on informational websites, where people want to find the information they are looking for in a fast and easy way.

4 Development of the Website Evaluation Questionnaire (WEQ)

The first version of the WEQ was tested on several municipal websites and on two websites that provide information but also entertain women (Cosmopolitan) and boys (Kaboem). In five studies different versions of the questionnaire were tested by 1104 respondents. Table 1 presents an overview of these studies.

Table 1. Five studies with different versions of WEQ

Websites the questionnaire is tested on	Number of respondents
Study 1: Cosmopolitan	465
Study 2: Kaboem	264
Study 3: Municipal website A	40
Study 4: Municipal website B	187
Study 5: Municipal website study C	148
Total number of respondents	1.104

Our main focus of analysis was on the reliability of the dimensions of the WEQ. We determined the reliability by computing the Cronbach's Alpha of every dimension. We aspired to reach for each sub dimension reliability scores higher than .70. Questions causing low reliability were revised or removed. This process was complemented in study 3 by think-aloud protocols of 40 respondents who commented on the questionnaire. This feedback helped us to diagnose the questions that resulted in low reliability scores, which led to three considerations for changing questions.

The first consideration concerned the perspective in every question. To stimulate people to give their own opinions (instead of taking on a jury role and speak for others) the questions were explicitly formulated from the respondent's perspective, as in *I find this website easy to use* versus *This website is easy to use*.

A second consideration was the finding that it is difficult for people to handle negations. Results of think-aloud protocols showed that people found it difficult to disagree with a negatively formulated assertion. This effect seems stronger when the word 'not' is used than when the negative connotation is in the word itself, like in 'not useful' versus 'useless'. So we tried to avoid the word 'not' in the questions.

A third consideration was the use of jargon. Several words proved to be difficult for people and were not interpreted correctly. An example is the term 'structure' which obviously led to very different interpretations. Some respondents gave their opinion about the menu on the homepage, others judged the quality of the links or judged to what extent they got lost on the website. The present WEQ contains five questions about the structure of the website and in only one of them the word 'structure' is used. In this way we can see to what extent the answers on the explicit structure question correspond with the other questions.

A factor analysis was used in order to assess whether the dimensions we distinguished were confirmed by the data. Results showed that four sub dimensions did not appear to measure one distinct construct. The sub dimensions *accuracy* and *comprehensiveness* had to be combined into one *comprehensiveness* sub dimension, and the sub dimensions *comprehensibility* and *language use* were combined in a new *comprehensibility* sub dimension. This resulted in the structure shown in Figure 1.

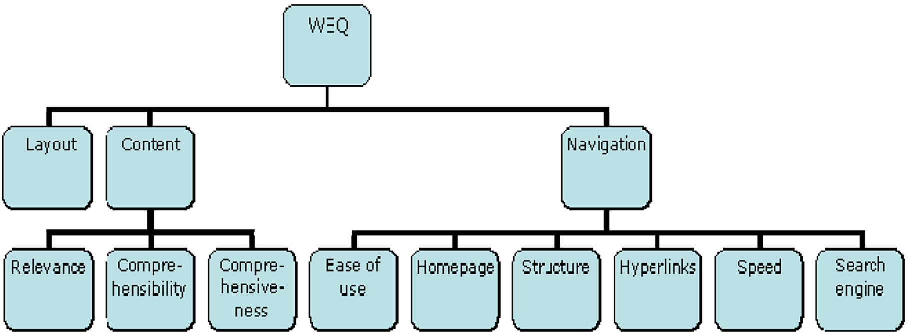

Fig. 1. Dimensional structure of the WEQ

Figure 1 presents the dimensional structure of the WEQ. The dimension of *Navigation* is related to attitudes towards the *process* of looking for information in the website. The dimension of *Content* is related to attitudes towards the *outcome* of this process: the information that is found in the website. Separate is the dimension of *Lay out* that is related to the so-called "look and feel" of the website. In the Appendix all questions that correspond to these dimensions are presented. In practice the questions on each dimension are not presented together, but are distributed throughout the questionnaire.

5 Assessment of the Reliability and Validity of the WEQ

This final version of the WEQ was evaluated in two studies. In the first study 408 respondents used the questionnaire to evaluate 18 municipal websites. The structure of the WEQ and the reliability were estimated by means of Linear Structural Relations (Lisrel). In the second study we tested the congruent validity of the WEQ; 19 participants performed two tasks on a municipal website, filled out the questionnaire afterwards and then commented on their scores.

With Lisrel we estimated the model in figure 1 with ten correlated factors. The correlations between the factors show the mutual coherence between the constructs. For example the correlation between *Relevance* and *Comprehensiveness* is .80, which means that these constructs measure different things, but also are also closely connected and in this case both measure an aspect of *Content*. The correlation between the dimensions *Homepage* and *Hyperlinks* is .97, which means that it is doubtful that these constructs measure different things. We have therefore decided to put them together in one dimension *Hyperlinks*.

The reliability is determined by means of Lisrel for the complete questionnaire and for the different dimensions and sub dimensions of the WEQ. Table 2 shows that the total reliability of the WEQ is high, with a score of .97. The dimensions *content*, *navigation* and *layout* also have very good scores of .88, .96 and .88 respectively. All sub dimensions, except *comprehensiveness,* have scores above .70. There are four sub dimensions marked with an asterisk. In these dimensions one question is removed to increase the reliability. In the Appendix, these questions are marked with an asterisk.

Table 2. Reliability scores of WEQ dimensions

Dimension	Number of items	Reliability
WEQ total	32	.97
Content	10	.88
Relevance	3	.72*
Comprehensibility	4	.75*
Comprehensiveness	3	.69
Navigation	19	.96
Ease of use	3	.90*
Structure	5	.80
Hyperlinks	6	.81*
Speed	2	.76
Search engine	3	.86
Layout	3	.88

* = one question removed

In a second study, we tested the congruent validity of the WEQ. We examined how attitude scores of respondents related to the experiences they had when visiting the website. We manipulated the tasks participants had to perform in such a way that one group was expected to have negative experiences in navigating a website and another group was expected to have positive experiences in the process of navigation. Both groups visited the same website, but with different tasks. Our hypothesis was that the first group would produce a negative attitude score on the items belonging to the dimension of *navigation* while the other group would produce a positive score on the items of this dimension. The same kind of manipulation was on the level of *content*: the first group with the difficult navigation task finally came across easy content, while the other group performing an easy navigation task was confronted with difficult content. After performing the tasks all participants (N=19) answered the questions presented in the WEQ. Afterwards they were asked to think aloud retrospectively while explaining their experiences on the website and their considerations when giving judgments on the questionnaire.

In order to assess the quality of our manipulation we scored the verbalizations of the participants while performing their tasks. An analysis of these scores confirmed that both groups had different experiences during navigation. Participants with a difficult navigation task needed significantly more time to perform their task than participants with an easy navigation task (13 minutes vs. 8 minutes, p < .05). They also made on average more negative comments about the navigation process than the participants with the easy navigation task (9,1 vs. 1,3; p < .05). There was also a difference in the mean number of comments on the content of the website between the group with a difficult and an easy comprehension task (3,1 versus 1,3; p <.05). Thus, we may conclude that participants indeed experienced different processes while navigating and comprehending the information.

In order to assess the validity of the WEQ we then analyzed the scores of both groups on the items of the dimensions of *navigation* and *content*. There was a significant difference between the groups on the sub dimension *hyperlinks* with scores from 3.2 (difficult navigation) and 4.0 (easy navigation) on a five point scale (p<.05). There were no significant differences on the other (sub) dimensions. The mean scores on *navigation* were rather positive, ranging from 3.4 (difficult navigation) to 3.7 (easy navigation) on a five point scale. The mean scores on the dimension *content* were even more positive: 4.0 (difficult content) versus 4.1 (easy content).

After having filled out the questionnaire, all participants commented on their scores. The analysis of this feedback provided several explanations for the observation that attitude scores were more positive than what would be expected considering the experiences respondents had while visiting the website.

First, people seem to focus stronger on the final result than on the process when thinking about a website. When people had found the information they were looking for, their attitude towards the process seemed to be overruled by the positive experience of finding and comprehending the information. In the protocols we often found statements such as: "I gave this positive score because I have found the information I needed." They seem to forget the complaints they had earlier in the process, when they had no idea where to go to.

A second explanation for unexpected positive attitude scores is that respondents often blamed themselves for problems they experienced. Respondents said that they had problems with reading texts, that they just did not think logically or that they always have problems finding information on the internet. They assigned their blame not to the designers of the website but to themselves, like Schriver [13] and Serenko [14] also reported in the context of difficulties with consumer electronic products and interface agents, respectively.

A third explanation can be found in the benchmark respondents use while expressing their attitudes towards the website. Some of the respondents told the evaluator that all government websites are boring. They do not expect to have an easy navigation process and to find information that is easy to understand. This leads to a low standard against which the website is judged. Negative experiences may result in positive attitudes because elsewhere respondents may have had considerably more trouble finding the right information.

6 Discussion

The WEQ appears to be a useful instrument to evaluate municipal and other governmental websites. The nine dimensions measure the attitudes of respondents about the navigation and the quality of the information in a reliable way. It is important that governmental organizations can use this standard questionnaire for evaluating their websites and for benchmarking their results against each other.

Research has shown, however, that we need to be careful in interpreting the results of the questionnaire. Respondents tend to give more positive attitude scores than what would be expected considering the experiences they have during visiting a website. Reasons for this are that respondents have a tendency of blaming themselves and of benchmarking against other websites. When interpreting the results this positive tendency has to be taken into account. This tendency is strongest in attitude scores about navigation. Scores about the process can change when respondents have found the information and have a positive attitude about the end result.

A subject that requires our permanent attention is the user friendliness of the WEQ. To keep respondents motivated, the WEQ should not be too long, should only consist relevant questions and the feeling of repetition should be kept down to a minimum. At the same time there is the concern of a good reliability and the diagnostic value of the WEQ. In the future we will more actively use the *routing*, which means that we leave out questions that are not relevant for users. For example the questions about the search engine will only be presented if respondents used this to search for information. In this way we try to create a questionnaire that is of high quality and is user friendly at the same time.

Acknowledgement

This paper is based on a research project that was financed by the Dutch Organization for Scientific Research (NWO). It is part of the research program Evaluation of Municipal Websites. The authors want to thank B. van Kippersluis, F. Loonstra, D. van Rijswijk and the students who participated in the master class Web Evaluation for their contributions to the development of the WEQ. We thank R. Supheert and her students for translating the WEQ from Dutch into English.

References

1. Spyridakis, J.H., Wei, C., Barrick, J., Cuddihy, E., Maust, B.: Internet-based research. Providing a foundation for web-design guidelines. IEEE Transactions on Professional Communication 48, 242–260 (2005)
2. Nielsen, J., Loranger, H.: Prioritizing Web usability. New Riders, Berkeley (2006)
3. ISO: Ergonomic requirements for office work with visual display terminals (VDTs)-Part 11: Guidance on usability (ISO 9241-11) (1998)
4. Hornbæk, K.: Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies 64, 79–102 (2006)
5. Sudman, S., Bradburn, N., Schwarz, N.: Thinking about answers: The application of cognitive processes to survey methodology. Jossey-Bass, San Francisco (1995)
6. Couper, M.P.: Web surveys. A review of issues and approaches. Public Opinion Quarterly 64, 464–494 (2000)

7. Dillman, D.A., Bowker, D.K.: The Web questionnaire challenge to survey methodologists. In: Reips, U., Bosjnak, M. (eds.) Dimensions of Internet science, pp. 159–178. Pabst Science Publishers, Lengerich (2001)
8. Molenaar, N.: Response effects of 'formal' characteristics of questions. In: Dijkstra, W., Van der Zouwen, J. (eds.) Response behaviour in the survey-interview, pp. 49–89. Academic Press, London (1982)
9. Kirakowski, J., Claridge, N., Whitehand, R.: Human-centered measures of success in Web site design. In: Proceedings of the 4th Conference on Human Factors & the Web, Baskerville, NJ, pp. 1–9 (1998)
10. Van Schaik, P., Ling, J.: Five psychometric scales for online measurement of the quality of human-computer interaction in Web sites. International Journal of Human-Computer Interaction 18, 309–322 (2005)
11. Lavie, T., Tractinsky, N.: Assessing dimensions of perceived visual aesthetics of web sites. International Journal of Human-Computer Studies 60, 269–298 (2004)
12. Muylle, S., Moenaert, R., Despontin, M.: The conceptualization and empirical validation of web site user satisfaction. Information & Management 41, 543–560 (2004)
13. Schriver, K.A.: Dynamics in document design. John Wiley, New York (1997)
14. Serenko, A.: Are interface agents scapegoats? Attributions of responsibility in human-agent interaction. Interacting with computers 19, 293–303 (2007)

Appendix: The Website Evaluation Questionnaire (WEQ)

Relevance

*I find the information in this website helpful.
The information in this website is of little use to me.
This website offers information that I find useful.

Comprehensibility

*I think the information in this website is described clearly.
The language used in this website is easy to me.
I find the information in this website easy to understand.
I find many words in this website difficult to understand.

Comprehensiveness

Certain information I was looking for was missing in this website.
The website provides me with sufficient information.
I find the information in this website precise.

User friendliness

I find this website easy to use.
*I had difficulty using this website.
I consider this website user friendly.

Structure

I know where to find the information I need on this website.
I was constantly being redirected on this website while I was looking for information.
I always know where I am on this website.

I find the structure of this website clear.
The convenient set-up of the website helps me find the information I am looking for.

Hyperlinks (including Homepage)
The homepage clearly directs me towards the information I need.
The homepage immediately points me to the information I need.
*I find the homepage confusing.
*I think it is difficult to spot the hyperlinks on this website.
It is clear which hyperlink will lead to the information I am looking for.
Under the hyperlinks, I found the information I expected to find there.

Speed
I think it takes a long time to download a new web page from this site.
I think this is a fast website.

Search Option
The search option on this website helps me to find the right information quickly.
The search option on this website gives me useful results.
The search option on this website gives me too many irrelevant results.

Layout
I think this website looks unattractive.
I like the way this website looks.
I find the design of this website appealing.

Translated from Dutch. Respondents can give their reactions to these assertions on five-point Likert scales (strongly disagree, disagree, neutral, agree, strongly agree).