Government Information Quarterly 29 (2012) 383–393

Contents lists available at SciVerse ScienceDirect

Government Information Quarterly

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

Measuring the quality of governmental websites in a controlled versus an online setting with the 'Website Evaluation Questionnaire'

Sanne Elling [a,*], Leo Lentz [a], Menno de Jong [b], Huub van den Bergh [a]

[a] Utrecht Institute of Linguistics (UiL-OTS), Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands
[b] University of Twente, Faculty of Behavioral Sciences, Department of Technical and Professional Communication, P.O. Box 217, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Available online 11 May 2012

Keywords:
Governmental websites
Usability
Questionnaires
Website quality
Multidimensionality

ABSTRACT

The quality of governmental websites is often measured with questionnaires that ask users for their opinions on various aspects of the website. This article presents the Website Evaluation Questionnaire (WEQ), which was specifically designed for the evaluation of governmental websites. The multidimensional structure of the WEQ was tested in a controlled laboratory setting and in an online real-life setting. In two studies we analyzed the underlying factor structure, the stability and reliability of this structure, and the sensitivity of the WEQ to quality differences between websites. The WEQ proved to be a valid and reliable instrument with seven clearly distinct dimensions. In the online setting higher correlations were found between the seven dimensions than in the laboratory setting, and the WEQ was less sensitive to differences between websites. Two possible explanations for this result are the divergent activities of online users on the website and the less attentive way in which these users filled out the questionnaire. We advise to relate online survey evaluations more strongly to the actual behavior of website users, for example, by including server log data in the analysis.

1. Introduction

The need to evaluate the quality of governmental websites is widely acknowledged (Bertot & Jaeger, 2008; Loukis, Xenakis, & Charalabidis, 2010; Van Deursen & Van Dijk, 2009; Van Dijk, Pieterson, Van Deursen, & Ebbers, 2007; Verdegem & Verleye, 2009; Welle Donker-Kuijer, De Jong, & Lentz, 2010). Many different evaluation methods may be used, varying from specific e-government quality models (e.g., Loukis et al., 2010; Magoutas, Halaris, & Mentzas, 2007) to more generic usability methods originating from fields such as human–computer interaction and document design. These more generic methods can be divided into expert-focused and user-focused methods (Schriver, 1989). Expert-focused methods, such as scenario evaluation (De Jong & Lentz, 2006) and heuristic evaluation (Welle Donker-Kuijer et al., 2010), rely on the quality judgments of communication or subject-matter experts. User-focused methods try to collect relevant data among (potential) users of the website. Examples of user-focused approaches are think-aloud usability testing (Elling, Lentz, & de Jong, 2011; Van den Haak, De Jong, & Schellens, 2007, 2009), user page reviews (Elling, Lentz, & de Jong, 2012), and user surveys (Ozok, 2008). In the *Handbook of Human–Computer Interaction* the survey is considered to be one of the most common and effective user-focused evaluation methods in human–computer interaction contexts (Ozok, 2008). Indeed, many governmental organizations use surveys to collect feedback from their users and in this way assess the quality of their websites. Three possible functions of a survey evaluation are providing an indication and diagnosis of problems on the website, benchmarking between websites, and providing post-test ratings after an evaluation procedure. A survey is an efficient evaluation method, as it can be used for gathering web users' opinions in a cheap, fast, and easy way. This, however, does not mean that survey evaluation of websites is unproblematic. The quality of surveys on the Internet varies widely (Couper, 2000; Couper & Miller, 2008). Many questionnaires seem to miss a solid statistical basis and a justification of the choice of quality dimensions and questions (Hornbæk, 2006). In this paper we present the Website Evaluation Questionnaire (WEQ). This questionnaire can be used for the evaluation of governmental and other informational websites. We investigated the validity and the reliability of the WEQ in two studies: the first in a controlled laboratory setting, and the second in a real-life online setting. Before we discuss the research questions and the design and results of the two studies, we will first give an overview of issues related to measuring website quality and discuss five questionnaires on website evaluation.

1.1. Laboratory and online settings

Surveys for evaluating the quality of websites can be administered in several different situations and formats. Traditionally, survey questions were answered face-to-face or with paper-and pencil based surveys, which needed to be physically distributed, filled out, returned, and

* Corresponding author.
E-mail addresses: s.elling@uu.nl (S. Elling), l.r.lentz@uu.nl (L. Lentz), m.d.t.dejong@utwente.nl (M. de Jong), h.vandenbergh@uu.nl (H. van den Bergh).

0740-624X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.giq.2011.11.004

then manually processed and analyzed. Currently, most surveys are filled out on a computer and data processing is automated (Tullis & Albert, 2008). The context in which computer-based surveys are used varies from a controlled situation in a usability laboratory, to an online real-life situation in which self-selected respondents visit a website in their own environment.

In an online setting, users with all kinds of backgrounds visit a website with a range of different goals. During their visits they navigate to various pages, using different routes to reach the information, for answering varying questions. Consequently, when the website is evaluated using the survey all respondents base their opinions on different experiences on the same website.

In a laboratory setting, participants generally conduct a series of tasks using a website, often in combination with other evaluation approaches such as thinking aloud, and fill out a questionnaire afterwards. The survey can be presented to participants after each task, the so-called *post-task ratings*, or after completion of a series of tasks as a *post-session rating* (Tullis & Albert, 2008).

An intermediate form between the laboratory and the online setting is a *remote evaluation* in which the questionnaire is filled out by a selected group of respondents who are invited to participate and who can choose their own time and place to do so. In some respects this remote evaluation resembles an online setting, but in other respects it resembles a laboratory setting.

The setting of the survey evaluation affects four methodological issues: the extent of control over the domain of evaluation, the ecological validity, the selection of respondents, and the accuracy of the answers.

The first issue is the control over the experiences on which respondents base their judgments expressed in the questionnaire. In an online setting, respondents may have a range of different goals; they may have visited hundreds of different pages or just two, reached by different navigation routes, some by clicking links, and some by using a search engine. Governmental websites often cover an extensive amount of web pages with information about a wide range of topics. This means that some respondents base their opinions on information related to obtaining a copy of a birth certificate, others on information about opening hours of the local swimming pool, or others on public transport information. So, like many other computer systems, the website can be seen as a compilation of components and not as one single entity (Brinkman, Haakma, & Bouwhuis, 2009). Moreover, some people fill out the questionnaire right at the beginning of a session, based on former experiences or on their first impressions, while others may fill it out after spending some time on the website. These divergent experiences make it difficult to measure user opinions validly and reliably in an online setting. Also, the interpretation of the answers in the questionnaire and the diagnosis of problems are more problematic because of the large range of underlying experiences. In a laboratory setting, the scope of the tasks is limited to a specific part of the website. An advantage of the laboratory setting is that evaluators know exactly which tasks have been conducted in which order, so there is no doubt on which parts of the website the judgments are based. Moreover, it is clear that respondents first completed their tasks and filled out the questionnaire afterwards, expressing judgments based on their experiences during task performance. This facilitates the comparison of user opinions and the diagnosis of the problems they encounter. To sum up, in an online setting respondents base their judgments on a very diverse set of experiences, while in a laboratory setting there is more control over and uniformity in the tasks respondents perform before filling out the questionnaire.

The second issue involves the ecological validity of the evaluation. Online respondents work in a natural situation. They are looking for information they choose themselves and they consider relevant. This is different from a laboratory setting, in which respondents usually work on predefined scenario tasks. These tasks are often made as realistic as possible, but will always remain artificial to some extent. Other confounding factors are the presence of a facilitator and the combination of task performance and evaluation. As a result, an online evaluation is more realistic than an evaluation in a laboratory setting.

The third issue involves the respondents who fill out the questionnaire. In a laboratory setting, the group of participants can be selected by the evaluator, who can try to draw a representative sample from the target population. This selection is expensive and time consuming, so the number of participants is often limited and the sample will not always be perfectly representative. The advantage of an online evaluation is that large numbers of participants can be reached. In principle all visitors of a website have the chance to share their opinions about this website. The selection is much cheaper and easier than in a laboratory setting. However, the self selection of high numbers of respondents also results in less control and a higher risk of a respondent group that is not representative of the target population. Couper (2000) discusses four representativeness issues, two of which are relevant in this context. The *sampling error* refers to the problem that not all users have the same chance of participating in the survey. When, for example, the survey is only announced on the homepage it will be missed by users who enter the website via a search engine. The *nonresponse error* means that not every user wants to participate in the survey. Several reasons may prevent users from participating, such as a lack of interest or time, technical problems or concerns about privacy. This nonresponse error is an important issue that must be taken into account in online survey measurements. An overview of the factors affecting response rate is given by Fan and Yan (2010). They distinguish four stages in the web survey process, which include survey development, survey delivery, survey completion, and survey return. A variety of factors is discussed that might influence the response rate in each stage of the process, such as the length of the survey, incentives, and the use of survey software. In all, both in a laboratory and in an online setting, problems with representativeness may occur. However, because of the self selection of respondents, the risk of errors is larger in an online setting.

The fourth issue concerns the accuracy of the answers. When answering a survey question, respondents must understand the item, retrieve the relevant information, use that information to make the required judgments, and map their opinions to the corresponding scale (Tourangeau, Rips, & Rasinski, 2000; Tourangeau, 2003). In a laboratory setting, participants fill out the questionnaire in a designated time and under the supervision of a facilitator. This means that respondents may be more careful and precise and take more time for the response process than respondents who fill out the questionnaire at home. Online data seem to have a greater risk of inadequacy, and therefore answers may be of lower quality (Heerwegh & Loosveldt, 2008). Research on online surveys by Galesic and Bosjnak (2009) has shown that respondents provide lower quality answers at the end of a questionnaire. Also, survey break off occurs more frequently in online settings (Peytchev, 2009), whereas laboratory respondents in general finish their survey as asked by the supervisor. However, the laboratory setting may have drawbacks as well. In the laboratory the questionnaire is often combined with other evaluation measurements. Consequently, the time between task completion and the answering process may be longer, which might complicate the retrieval process. To conclude, both settings have aspects that may threaten the accuracy of the answers, but the risks seem higher in online settings.

In sum, on the one hand it is useful to measure the opinions of people who are using a website in natural settings, and who base their judgments on their own experiences (Spyridakis, Wei, Barrick, Cuddihy, & Maust, 2005). On the other hand, online settings have several risks which complicate the measurements. This raises the question whether the same questionnaire can be used in both an online and a laboratory setting, and whether the results of different evaluations can be compared without analyzing the effects of the settings on the measurement of the constructs, as is often done in practice.

1.2. Research on other questionnaires on website evaluation

Below we will discuss five questionnaires that can be used for measuring website quality: (1) the System Usability Scale (SUS) (Brooke, 1996), (2) the American Customer Satisfaction Index (ACSI) (Anderson & Fornell, 2000), (3) the Website Analysis Measurement Inventory (WAMMI), (Kirakowski, Claridge, & Whitehand, 1998), (4) a five-scale questionnaire (Van Schaik & Ling, 2005), and (5) the Website User Satisfaction Questionnaire (Muylle, Moenaert, & Despontin, 2004). These five questionnaires are prominent examples of usability questionnaires, the first three because they are often mentioned in the usability literature, and the other two because they have been comprehensively validated. We realize that many other questionnaires exist, but we chose to leave these aside because they are mentioned less often in the literature or are less well validated. Examples are the After Scenario Questionnaire (Lewis, 1991), the Expectation Measure (Albert & Dixon, 2003), the Usability Magnitude Estimation (McGee, 2004), the Subjective Mental Effort Questionnaire (Zijlstra, 1993), and several questionnaires that are discussed in Sauro and Dumas (2009), and Tedesco and Tullis (2006).

The five questionnaires in this overview are compared on six aspects. First, it is important that the questionnaire is available for general use and in this way open for analyses to assess their quality. Second, it should be clear to which domain the questionnaire applies. In this article we focus on informational governmental websites and we will therefore examine the usefulness of the five questionnaires for this domain. A third aspect is the function of the questionnaire: can it be used for diagnosing, benchmarking, and/or post-test ratings? Fourth, a questionnaire for measuring website quality should have some clearly defined dimensions that measure relevant aspects of quality. To determine this multidimensionality, the proposed factor structure should be tested against sample data to demonstrate whether the factor structure is confirmed and how the factors are related to each other. Fifth, it is important that quality aspects are measured reliably, which means that a scale should consistently reflect the construct that it is measuring. Sixth, these factors should be sensitive to differences between tested websites.

Many usability questionnaires are designed with the purpose to keep evaluations simple and cost-effective. These questionnaires are rather short, can be applied to a range of contexts and systems, and provide a general indication of the overall level of usability. An often used questionnaire is the System Usability Scale (SUS) (Brooke, 1996). This questionnaire consists of ten items (alternating positive and negative) on which respondents can indicate their level of agreement on five-point Likert scales. In the SUS, two dimensions can be distinguished, *usability* with eight items and *learnability* with two items (Lewis & Sauro, 2009). The questionnaire can be used for global quality assessment, global benchmarking, or as a post-test rating. The result of the evaluation is an overall SUS score between 0 and 100, which can be benchmarked against the scores of other systems. Bangor, Kortum, and Miller (2009) complemented the SUS with an eleventh question which measures the overall opinion about the system's user-friendliness. They used this score to put labels on the SUS-scores, so that these scores can be converted to absolute usability scores that can be interpreted more easily. In studies by Bangor, Kortum, and Miller (2008) and Lewis and Sauro (2009), the SUS had high reliability estimates and proved to be useful for a wide range of interface types. Tullis and Stetson (2004) compared the SUS with four other surveys and found that the SUS was best able to predict significant differences between two sites, even with small sample sizes. However, the short and simple design of the SUS and the wide range of interfaces it can be applied to may also have their drawbacks. When the SUS is used for the evaluation of an informational website it will only give a very general impression of its quality with limited diagnostic value. Moreover, it is questionable whether the ten items which are applicable to so many interfaces really represent the most salient quality features of an informational website.

Another frequently used questionnaire is the American Customer Satisfaction Index (ACSI) by Anderson and Fornell (2000), aimed at measuring quality and benchmarking between websites. This questionnaire also measures user satisfaction in a wide range of contexts. However, the ACSI contains questions that can be adjusted to the function of the website. For informational websites the questionnaire consists of the elements *content*, *functionality*, *look and feel*, *navigation*, *search*, and *site performance*. All these elements contribute to an overall user satisfaction score which can be compared to other websites' scores. It is unclear how these questions really apply to informational websites, as the same questions seem to be used for private sector sites such as online news sites and travel sites. Also comparisons between online and offline government services have been made with the ACSI. How exactly the ACSI is constructed and to what extent comparisons between websites and services are based on the same questions, has not been reported. Measurements of reliability or validity have not been made public, so it is difficult to judge the quality of this questionnaire and to compare it to others.

A third questionnaire that is often mentioned in usability contexts is the Website Analysis Measurement Inventory (WAMMI) by Kirakowski et al. (1998). The WAMMI is composed of 20 questions (stated positively or negatively), which have to be answered on five-point Likert scales. The questions are divided into five dimensions: *attractiveness*, *controllability*, *efficiency*, *helpfulness*, and *learnability*. Kirakowski et al. reports high reliability estimates, between 0.70 and 0.90, for the five dimensions. However, these estimates are computed for a version of the WAMMI that consisted of 60 questions. It is unclear to what extent these same high estimates are achieved in the 20 question version that is used in practice. The fact that the WAMMI is frequently used, offers the advantage that scores can be compared against a reference database with tests of hundreds of sites, which makes it suitable for benchmarking. A limitation of this questionnaire however, is the limited justification of reliability and validity issues.

Fourth, there is a questionnaire compiled by Van Schaik and Ling (2005), consisting of five scales for the online measurement of website quality. This questionnaire was validated more extensively than the first three questionnaires we discussed. The dimensions of this questionnaire are: *perceived ease of use*, *disorientation*, *flow* (*involvement and control*), *perceived usefulness*, and *esthetic quality*. Van Schaik and Ling investigated the sensitivity of the psychometric scales to differences in text presentation (font) on a university website. Students performed retrieval tasks and filled out the questionnaire afterwards. The factor analysis revealed six distinct factors, as *flow* fell apart into two separate factors (*involvement* and *control*). All factors had high reliability estimates, ranging from 0.74 to 0.97 (based on three to seven questions for each factor). No effects of font type were found on the six dimensions, so in this study the questionnaire was not sensitive to differences in text presentation on websites. The authors expect that stronger manipulations of text parameters will demonstrate the validity and sensitivity of scales more clearly. Their research was only administered with students; it would be useful to also test the questionnaire with respondents with different educational backgrounds, experience, and age.

Another well-founded questionnaire is the fifth and last we discuss: the Website User Satisfaction Questionnaire by Muylle et al. (2004). This questionnaire was developed for the evaluation of commercial websites. It was based on theories about hypermedia design and interactive software and on a content analysis of think-aloud protocols aimed at eliciting relevant dimensions of website user satisfaction. In this way a 60-item questionnaire was developed and tested with a sample of 837 website users who filled out the questionnaire after performing tasks on a website of their own choice. A confirmatory factor analysis supported the distinction in four main dimensions and eleven sub dimensions. The first dimension is *connection* with the sub

dimensions *ease of use*, *entry guidance*, *structure*, *hyperlink connotation*, and *speed*. The second dimension is the *quality of information* with the sub dimensions *relevance*, *accuracy*, *comprehensibility*, and *completeness*. The third and fourth dimensions are *layout* and *language*, which do not have sub dimensions. In their study, Muylle et al. used 60 items, 26 of which were dropped afterwards based on correlations and reliability estimates. The dimensions have high reliability estimates, between .74 and .89. It remains uncertain, however, to what extent the same estimates would be obtained if the 34-item questionnaire would be tested. The dimensions represent clearly defined aspects of website quality, which results in an adequate diagnostic value of the questionnaire. However, there is no information about the extent to which the questionnaire is able to show differences between websites.

The SUS, the ACSI, and the WAMMI are mentioned most frequently in the usability literature. However, they do not seem to be based on a profound analysis of validity and reliability issues. The wide range of contexts they can be used in, raises doubts about the suitability of these questionnaires for an informational, governmental website context. The questionnaires by Van Schaik and Ling (2005) and Muylle et al. (2004) are more extensively validated but appear to be absent in usability handbooks and in usability practice. These two questionnaires are not specifically designed for informational websites. Van Schaik and Ling involve dimensions in their questionnaire that are less relevant in an informational governmental context, such as *flow*, and Muylle et al. explicitly focus on commercial websites.

In conclusion, we can say that a well-founded questionnaire for the domain of informational governmental websites is not available yet. We therefore developed the Website Evaluation Questionnaire (WEQ), which will be described below.

1.3. The Website Evaluation Questionnaire (WEQ)

The WEQ focuses on the domain of governmental websites. This questionnaire can be used for detecting and diagnosing usability problems, for benchmarking governmental websites, and as a post-test rating. The questionnaire may also be suitable for other kinds of informational websites that have the primary aim to provide knowledge to users without commercial or entertainment motives. To enable users to find answers on their questions efficiently on these websites, three main aspects are important. First, the information should be easy to find. Second, the content should be easy to understand. Third, the layout should be clear and should support users' adequate task performance. Consequently, website quality splits into several components and should be measured with different questions, which are spread over several relevant dimensions.

The WEQ was developed on the basis of literature on usability and user satisfaction. Muylle's et al. (2004) questionnaire was used as the main source, complemented by other theories. After several evaluations the WEQ was refined to the version presented in this article. An elaborate description of this development process can be found in Elling, Lentz, and De Jong (2007). The WEQ evaluates the quality of the three relevant aspects of governmental websites described above. The dimension *navigation* measures the opinions of users on the information seeking process. The dimension *content* measures the outcome of this process: the quality of the information found on the website. Both dimensions are composed of various sub-dimensions which are shown in Fig. 1. The third dimension is *layout*, which is related to the so-called “look and feel” of the website. The complete questionnaire can be found in Appendix A.

To what extent is the multidimensional structure presented in Fig. 1 confirmed by evaluation data? In some preliminary studies, described in Elling et al. (2007), the WEQ was tested in several contexts, and its reliability and validity were evaluated. The results showed that both validity and reliability were satisfactory, but also called for some adjustments on a global level as well as on more detailed levels of question wording. The current study uses the new version of the WEQ as a starting point and addresses the psychometric properties of the WEQ in controlled and online settings.

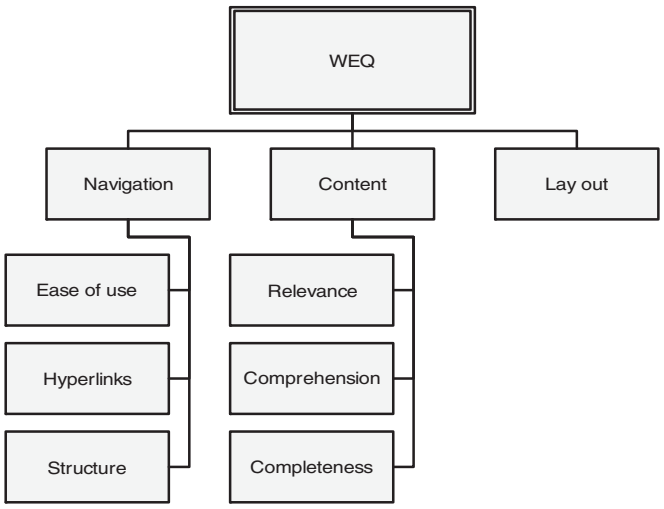

Fig. 1. Multidimensional structure of the Website Evaluation Questionnaire for governmental websites.

1.4. Research questions

We applied the WEQ in two separate studies: in a controlled laboratory setting and in an online setting. In the laboratory setting, participants performed tasks on three websites and filled out the questionnaire afterwards. In the online setting the questionnaire was placed on four governmental websites. The main research question is: can the multidimensional structure of the WEQ be justified in the controlled setting and to what extent is it confirmed in online settings?

1.4.1. Psychometric properties WEQ in a controlled setting

First, we will focus on the WEQ in a controlled setting. Does the questionnaire have clearly distinguishable factors which each measure different aspects of website quality? Results of the questionnaire can only be interpreted and diagnosed in a meaningful way if it measures the same constructs across different websites. This means that the latent multidimensional structure of the WEQ should be consistent for different websites. Only if the questionnaire measures the same constructs on different websites, can it be used to uncover quality differences between these sites and for benchmarking between them. So, the first research question is:

- Does the WEQ have a demonstrable factor structure in which multiple dimensions can be distinguished and which is consistent for different governmental websites?

Second, we will investigate whether the distinct factors measure user opinions reliably. The reliability for a set of questions examining the same construct is a measure for the proportion of systematic variance as compared to the proportion of error variance. A high reliability estimate means that the proportion of systematic variance is large. This leads to the second question:

- To what extent do the dimensions of the WEQ measure website quality aspects reliably?

If the factor structure is indeed consistent and reliable, the WEQ should be sensitive to differences between the websites. After all, one of the purposes of an evaluation is often to identify quality differences between websites. This leads to the third research question:

- To what extent does the WEQ discriminate between different governmental websites?

If the WEQ shows adequate psychometric properties in the controlled setting, we can switch to an online setting and test the validity and the reliability in this more complex situation, with a variety of experiences and risks of inadequacy.

1.4.2. Psychometric properties WEQ in an online setting

We will start with comparing the factor structure of the laboratory setting and the online setting, using multiple group confirmatory analysis to check if the multidimensional structure of the WEQ is consistent across laboratory and online settings. So, the first research question in the online setting is:

- To what extent is the WEQ consistent in laboratory and online settings?

Then, we will answer the same three research questions we used in the laboratory setting, by measuring the stability of the factor structure over four governmental websites, the reliability of the dimensions, and the sensitivity of the WEQ to differences between the four websites.

2. WEQ in controlled settings (study 1)

2.1. Method

2.1.1. Respondents

The WEQ was tested in laboratory settings on three governmental websites. In total, 273 participants took part in the laboratory studies: 90 participants for each of the first two websites, and 93 for the third website. All respondents were selected by a specialized agency that maintains an extensive database of potential research participants. The participants received financial compensation for taking part in the study.

All participants indicated they used the internet at least once a week. The selection of the sample was based on gender, age, and educational level, following the criteria proposed by Van Deursen and Van Dijk (2009). Men and women were almost equally represented with 130 (48%) males and 143 (52%) females. Participants were divided into four different age categories: 18–29 (62 participants, 23%), 30–39 (65 participants, 24%), 40–54 (76 participants, 28%), and 55 and older (70 participants, 26%), which were divided equally over each website. There were three educational levels (low, medium, and high), based on the highest form of education people had received. The group with the lowest education level ranged from elementary school to junior general secondary professional education (79 participants, 29%). The group with the medium education level had intermediate vocational education, senior general secondary education or pre-university education (91 participants, 33%). The highly-educated group consisted of higher vocational education or university level participants (103 participants, 38%). All groups were divided equally over the three websites. All characteristics were mixed in such a way that, for example, all age categories and genders were equally spread over all educational levels.

2.1.2. Procedure

In the controlled setting participants filled out the questionnaire on a computer, after finishing two or three scenario tasks on one of the three governmental websites. They filled out the questionnaire at the end of the session, which means after task completion and other evaluation measurements.[1]

2.1.3. Material

The questionnaire was used to evaluate three governmental websites of medium to large Dutch municipalities. A municipal website is intended specifically for inhabitants, but sometimes also for tourists and businesses. These websites contain a variety of information, as they are designed to satisfy the informational needs of a broad target audience.

2.1.4. Analysis

To answer the first research question on the factor structure, the multidimensional structure of the WEQ was tested in a confirmatory factor analysis. To test the stability of the factor structure over websites, we did a cross validation on samples. So we analyzed the latent structure in different samples simultaneously. This was done by means of multiple group confirmatory factor analysis, with the use of Lisrel 8.71 (Jöreskog & Sörbom, 2001). With multiple group confirmatory analysis a hypothesized factor structure can be tested in different populations simultaneously, and in this way measurement invariance can be identified. The factor structure was tested with four nested models, each posing varying constraints on the measurement invariance between websites (Jöreskog, 1971). The *parallel model* is the most restrictive model. In this model it is assumed that all items load on the intended constructs with an equal reliability on all websites. That is, next to the invariance of the factor structure (equality of the correlations between factors) it is assumed that both true-score and error-score variance do not differ between websites. The *tau-equivalent model* allows for the possibility of different amounts of error-score variance. The *congeneric model* is the least restrictive, assuming that individual items measure the same latent variable but possibly with different amounts of true-score variance and error-score variance. The *non generic model* relinquishes the assumption that the same constructs are measured in different samples. The fit of these models can be tested by means of a chi-square distributed testing statistic, and be evaluated by other fit indices. Rijkeboer and Van den Bergh (2006) used similar techniques for their research on a questionnaire for the assessment of personality disorders, the Young Schema-Questionnaire (Young, 1994). They provide an elaborate overview of the literature about these techniques.

Lisrel was also used to answer the second research question. The reliability estimates of the seven dimensions of the WEQ were tested, based on the principle that a scale should consistently reflect the construct it is measuring. A univariate general linear model (ANOVA) was used to answer the third research question and thus determine the extent to which the WEQ is able to discriminate between websites.

2.1.5. Indices of goodness of fit

We used several indices of goodness of fit to compare the latent structures of the WEQ across different samples. First, we looked at the chi-square differences between the four nested models of the factor structure to decide which of the models showed the best fit. However, chi-square is affected by sample size: a large sample can produce larger chi-squares that are more likely to be significant and thus might lead to an unjust rejection of the model. Therefore, also four other indices of goodness of fit were taken into account. These were firstly, the *critical N*: the largest number of participants for which the differences would not be significant and the model would be accepted. The second index is the *normed fit index* (*NFI*), which varies from 0 to 1 and reflects the percentage of variance that can be explained. The closer this index is to 1, the better the fit. Values below .90 indicate a need to adjust the model. The third index is the *comparative fit index* (*CFI*), which is also based on the percentage of variance that can be explained. Values close to 1 indicate a very good fit, values above .90 are considered acceptable. The fourth index is the *root mean square residual* (*RMR*), which shows the percentage

[1] Participants were divided over four laboratory conditions. In the first condition participants were asked to review the website with a software tool for collecting user comments (Elling et al., 2012). In the other three conditions participants carried out tasks on the website while their eye movements were recorded and they were asked to think aloud during the task completion (condition 2) or afterwards while looking at a recording of their actions (condition 3), see Elling, Lentz and de Jong (2011). An elaborate description of these think-aloud conditions is presented in Van den Haak, De Jong and Schellens (2007, 2009). In half of the retrospective recordings a gaze trail of the eye movements was added (condition 4). Analyses have shown that the WEQ's multidimensional structure is consistent in the four conditions ($\chi^2 = 97.66$; df = 116; p = 0.89).

that is not explained by the model. There is a good model fit if this score is less than or equal to .05 and the fit is adequate if the score is less than or equal to .08.

2.2. Results

Research question 1: To what extent does the WEQ have a demonstrable factor structure in which multiple dimensions can be distinguished and which is consistent for different websites?

Before differences in means between websites can be compared meaningfully we need to assess the fit of a model that shows that (1) the factor structure is consistent over websites, and (2) the seven dimensions can be distinguished empirically. In Table 1 three comparisons of different nested models are shown, testing the fit of four models.

Table 1 shows that the difference in fit of the congeneric model is significantly better than either the parallel or the tau-equivalent model. The highest row shows that tau-equivalent significantly fits better than parallel; the second row shows that congeneric significantly fits better than tau-equivalent. In other words, the congeneric model fits to the observed data better than either the parallel or the tau-equivalent model. The difference in fit between the congeneric and non-congeneric model however, proved to be non-significant (p = .89). Therefore, the congeneric model is the model that best fits the data. The absolute fit of this model can be described as adequate ($\chi^2 = 945.7$; df = 669; p = <.001; CFI = .97; NFI = .94; RMR = .06). Although the χ^2-testing statistic is somewhat high, the other statistics indicate a good fit of the model to the observed data. Therefore, we conclude that the factor structure of the WEQ is (1) consistent over websites, although (2) the reliability of the different dimensions fluctuates between websites (see also Table 3) and (3) that seven factors can be distinguished empirically (see also Table 2).

Table 2 shows the correlation matrix in which the correlations between the seven dimensions are reported. The correlations between the dimensions show that each dimension partly measures something unique. However, they are not completely different and do show some coherence. There seems to be no higher order structure, although the correlations between *ease of use*, *hyperlinks*, and *structure* are comparatively high. These three dimensions clearly represent an aspect of accessibility, which explains the higher correlations.

Research question 2: To what extent do the seven WEQ dimensions measure website quality reliably?

Table 3 shows the reliability estimates for the seven dimensions of the WEQ for each of the three websites, based on the congeneric model. On website 1, all dimensions have a reliability estimate above .70. Also, on the other two websites most dimensions are above .70, but both websites have two dimensions that are (a little) under .70: *comprehension* with estimates of .65 and .54, *structure* with an estimate of .63, and *relevance* with an estimate of .66. As we explained earlier, the congeneric model allows for varying reliability estimates. This means that a dimension can provide reliable measures on one website, but not so well on another website. Most dimensions have good reliability estimates on all websites or on the majority of the websites, only the dimension *comprehension* requires attention. In all, we can conclude that the WEQ proved to be a reliable instrument, with some reservations for the *comprehension* dimension.

Table 1
Three comparisons of four nested models testing the invariance of the factor structure of the WEQ over websites.

Comparison models	χ^2	df	p
1. Parallel measurements versus tau-equivalent measurements	127.9	46	.00
2. Tau-equivalent measurements versus congeneric measurements	165	46	.00
3. Congeneric measurements versus non-congeneric measurements	31.1	42	.89

Note: χ^2 = chi-square statistic; df = degrees of freedom; p = level of significance.

Table 2
Correlation matrix for the laboratory setting.

Dimension	1	2	3	4	5	6	7
1. Ease of use	1						
2. Hyperlinks	0.78	1					
3. Structure	0.76	0.80	1				
4. Relevance	0.38	0.39	0.37	1			
5. Comprehension	0.42	0.35	0.33	0.41	1		
6. Completeness	0.47	0.49	0.51	0.62	0.49	1	
7. Lay out	0.31	0.27	0.35	0.37	0.19	0.30	1

Research question 3: To what extent does the WEQ discriminate between websites?

Given the clearly defined dimensions, which prove to be reliable and consistent for the different websites, we can answer the third research question. For each dimension we measured differences between websites, which are shown in Table 4.

Table 4 shows that the WEQ is indeed sensitive to differences between the three governmental websites. The differences were most obvious on the three navigation dimensions: *ease of use*, *hyperlinks*, and *structure*. There are two possible explanations for this result. First, perhaps the three municipal websites in our study differed more strongly in the way they structured their information and presented this structure on the homepage with links, than in the content and in the lay out. On governmental websites the layout is often quite basic and functional, which may explain why no differences were found on the layout dimension. Second, users may have focused more on the process of finding the information than on the outcome of this process. This may have resulted in stronger opinions on accessibility issues, which could cause more differences between the scores of the websites on these dimensions.

2.3. Summary of findings study 1

We can conclude that, for the laboratory setting, the WEQ demonstrated a stable underlying multidimensional structure. The congeneric model shows the best fit, which means that the WEQ measures the same constructs on the three websites, but possibly with differences in reliability, error variance, and true score variance. The reliability estimates show that all dimensions are able to measure opinions in a reliable way. So we can conclude that the WEQ proved to be a valid and reliable instrument that can be used for

Table 3
Reliability estimates per dimension on the three websites.

Dimension	Website 1 (N = 90)	Website 2 (N = 90)	Website 3 (N = 93)
Ease of use	.88	.87	.83
Hyperlinks	.83	.79	.71
Structure	.71	.73	.63
Relevance	.91	.66	.75
Comprehension	.74	.65	.54
Completeness	.70	.71	.70
Lay out	.91	.79	.80

Table 4
Mean scores (standard deviation) and significant differences between mean scores for each website and each dimension in the laboratory setting.

Dimension	Website 1 (N = 90)	Website 2 (N = 90)	Website 3 (N = 93)	Differences mean scores
Ease of use	3.30 (.85)	3.13 (.89)	2.70 (.90)	F (2, 273) = 11.51, p<.01
Hyperlinks	3.11 (.83)	2.89 (.75)	2.49 (.79)	F (2, 273) = 17.06, p<.01
Structure	3.30 (.72)	3.08 (.74)	2.95 (.68)	F (2, 273) = 5.33, p<.01
Relevance	3.84 (.65)	3.81 (.65)	3.65 (.69)	F (2, 273) = 1.55; ns
Comprehension	3.86 (.60)	3.80 (.63)	3.65 (.63)	F (2, 273) = 2.90; ns
Completeness	3.72 (.56)	3.71 (.62)	3.44 (.64)	F (2, 273) = 5.53, p<.01
Lay out	3.41 (.92)	3.42 (.78)	3.31 (.80)	F (2, 273) = 0.47; ns

Note: Opinions were measured on a five-point scale, where 1 is most negative and 5 is most positive (negatively stated items were reversed).

evaluating the opinions of users about websites in a laboratory setting. In the laboratory setting the WEQ is sensitive to differences between websites. This means that the instrument can be used for benchmarking between websites, by comparing the results on the dimensions.

3. WEQ in online settings (study 2)

3.1. Method

3.1.1. Respondents

Overall, 1585 respondents started filling out the questionnaire on one of the four governmental websites, of whom 1394 respondents completed the whole questionnaire. This means that 191 respondents (12%) dropped out somewhere in the questionnaire. This percentage is consistent with the findings of Peytchev (2009), who reports break off rates between 9% and 34%. However, not all respondents who completed the survey seem to have done this seriously: 34 respondents (2.4%), who evidently failed to answer the questions seriously, were filtered out of the dataset. These respondents were identified by checking the number of questions that were answered using the same scale position. This is an indication of less-attentive respondents, since the questionnaire also contained negatively stated items that should be rated oppositely (in the WEQ eight out of 25 items, one on each dimension). We put the limit on 20 items out of 25. So if respondents used the same scale position 20 times or more, they were excluded from the sample. This reduction brings the total online sample to 1360 respondents.[2]

As stated earlier, in an online setting there is little control over the selection of a representative group of respondents. However, it is possible to ask for some demographic factors and so subsequently determine the characteristics of the sample. All age-groups seem reasonably well-represented: 10% of the participants were between 18 and 29 years of age, 14% between 30 and 39 years, 21% between 40 and 54, and 19% of the participants were older than 55.[3] Somewhat more males than females filled out the questionnaire: 57% males versus 42% females. No information about gender was available for 1% of the users. When looking at the educational level, the most remarkable outcome is the high proportion of highly educated respondents. More than half of the respondents (57%) reported to have a high educational level, 33% of the respondents had a middle educational level, and the group of lower educated people was rather small (9%). These outcomes can to a certain extent be explained by the statistics on Internet behavior of groups with different educational levels, which reveal that in 2009 80% of higher educated people visited a governmental website in the past three months, 61% of middle educated people, and only 33% of lower educated people.[4]

Table 5
Correlation matrix for the online results (* = significant deviation from the laboratory correlation).

Dimension	1	2	3	4	5	6	7
1. Ease of use	1						
2. Hyperlinks	0.76	1					
3. Structure	0.82	0.82	1				
4. Relevance	0.57*	0.49	0.53*	1			
5. Comprehension	0.47	0.38	0.44	0.56	1		
6. Completeness	0.65*	0.62	0.65	0.71	0.47	1	
7. Lay out	0.57*	0.46*	0.54*	0.46	0.38*	0.44	1

3.1.2. Procedure

In the online setting the questionnaire was placed on four municipal websites for a period of about five weeks. Three of these websites were also evaluated in the laboratory setting. The questionnaire was announced on the home pages of the websites and on news and contact pages. The survey evaluation was also mentioned in the local media, to draw the attention of the citizens on the questionnaire and to persuade users of the website to fill it out. Some gift vouchers were divided among the participants, to stimulate them to take part in the survey. The questionnaire could be filled out only one time from the same computer.

3.1.3. Analysis

We started with an analysis to compare the factor structure of the laboratory data and the online data in a cross validation on samples. This was analyzed using a multiple group confirmatory factor analysis, to identify measurement invariance. The models of the laboratory and the online data were compared using standardized residuals. With this model the residuals can be compared on the 'standard scale.' The often used criterion for good fit is that numbers higher or lower than 2 indicate there is no good model fit. After analyzing the invariance between laboratory and online conditions, the stability of the factor structure over the four online websites was measured in order to answer the second research question. For the third research question about reliability, Lisrel was used to estimate the reliability estimates of the seven dimensions of the questionnaire for the online samples. To answer the fourth research question, we assessed the sensitivity of the WEQ for differences between websites, using a univariate general linear model (ANOVA).

3.2. Results

Research question 1: To what extent is the WEQ consistent over the laboratory and online settings?

Before interpreting the online opinion scores we need to answer the question to what extent the same underlying factor structure as measured in the laboratory is demonstrable in the online data set. The correlations between the dimensions for the online condition are shown in Table 5.

As in the laboratory study, all dimensions measure something different but show some overall coherence. Again, the three dimensions *ease of use*, *hyperlinks*, and *structure* are relatively highly correlated

[2] We applied the same filter to the laboratory data, but no respondents were filtered out in that sample.

[3] For website 1 no information about age was collected, which explains the high number of missing values concerning age: 483 (36%).

[4] Data come from Statline, the electronic databank of Statistics Netherlands: www.statline.cbs.nl last visited on February 26th 2010.

with each other, as they are all aspects of accessibility. However, no higher order structure can be derived from this matrix.

The conclusion that the same factors are distinguishable in the controlled and the online settings does not automatically mean that these dimensions measure the same constructs. To what extent do these scores represent the same constructs as those we measured in the controlled setting? Is the factor structure of the WEQ consistent across the laboratory and online settings? In other words, does the congeneric model fit the observed data in both conditions? The chi-square shows a difference between the constructs in the laboratory and online conditions ($\chi^2 = 121.32$; df = 44; $p < .01$). However, the number of respondents was high, which may lead to an unjust rejection of the model. The other indices of goodness of fit seem to point to a reasonably adequate fit of the model. The critical N is 1137, so with less than 1137 participants the model would fit and the factor structure would be confirmed. The NFI of 0.99 and the CFI of 1.00 both indicate an extremely good fit of the model. However, the RMR of 0.09 shows that some of the observed (co)variance is not explained by the model. In sum, we must conclude that the constructs do not entirely measure the same things in both conditions.

A comparison of the two correlation matrices can help explain the differences between the conditions. All correlations between the dimensions are higher in the online condition than in the laboratory setting. The correlations marked with an asterisk (see Table 5) are significantly higher in the online condition. The dimension *layout* strongly correlates with four other dimensions. Other significantly higher correlations can be seen between *ease of use* and *relevance*, between *ease of use* and *completeness*, and between *structure* and *relevance*. Hence, in the online situation less distinction is made between the different aspects of the website quality than in the laboratory condition. This might be explained by the variation in tasks and goals in the online setting, which leads to a diffuse mix of experiences on which the judgments are based. A second explanation may be that the online users filled out the questionnaire in a more global way, because they took less time and care to fill out the questionnaire than the respondents in the laboratory. They possibly formed a general overall impression, which dominated their response behavior and resulted in less diverse scores per dimension and higher correlations.

So, the answer to the first research question is that the same factors can be distinguished in the laboratory and online settings. These factors partly measure the same constructs, although in the online setting the factors more strongly influence each other. The layout construct differed most in the two settings.

Research question 2: To what extent is the factor structure consistent for the four different websites in the online condition?

To adequately compare the scores of the four websites, the factor structure should be consistent over websites. The chi-square distribution indicates differences between the websites ($\chi^2 = 162.28$; df = 116; $p < .01$). However, this can be explained by the high number of 1360 respondents. The other indices of fit indicate a consistent factor structure for the four websites. The critical N is 1222, which means that with 1222 respondents or less the model would fit. The NFI is .98 and the CFI is .99, so both show a good fit of the model. The RMR is .03, which also indicates a good fit. Based on the indices, we can assume that the WEQ measures the same constructs on the four websites.

Research question 3: To what extent do the dimensions of the WEQ measure website quality reliably?

The reliability estimates of the dimensions in the online setting are shown in Table 6. Here also, almost all dimensions are above .70. Only *comprehension* scores a little lower on three of the four websites.

Table 6
Reliability estimates per dimension on the four websites.

Dimension	Website 1 (N = 468)	Website 2 (N = 185)	Website 3 (N = 100)	Website 4 (N = 607)
Ease of use	.86	.87	.84	.88
Hyperlinks	.83	.86	.84	.85
Structure	.82	.81	.80	.84
Relevance	.81	.77	.85	.76
Comprehension	.63	.66	.76	.64
Completeness	.76	.78	.85	.77
Lay out	.84	.83	.81	.84

As in study 1, the dimensions have different reliability estimates on different websites. However, these differences are generally rather small.

Research question 4: To what extent can the WEQ discriminate between the four websites?

Table 7 shows the mean scores and standard deviations on each dimension for the four websites.

Is the questionnaire sensitive to differences in quality on the seven dimensions? A linear mixed model analysis revealed only one significant difference between websites. On *hyperlinks* website 4 scores significantly lower than the other websites ($F(3, 1359) = 3.57$, $p < .05$). On all other dimensions no significant differences between websites were found. This result may be explained by the high diversity between goals, pages visited, and experiences of users who filled out the questionnaire. Differences between websites may be present in reality, but these are rendered invisible by the differences between things people did on the website.

3.3. Summary of findings study 2

In the online setting the WEQ consists of the same seven dimensions as we distinguished in the controlled situation. This means that the WEQ has a stable multidimensional structure that is upheld even in complex online measurements. However, the dimensions partly reflect something different in the controlled and online settings. In the online setting the dimensions correlate more highly with each other, which might be due to the diverse range of goals and pages users base their judgments on. This diversity might also explain why it is difficult to distinguish between websites in an online setting.

Table 7
Mean scores (standard deviation) for each website and each dimension in the online setting (* = significant difference between websites).

Dimension	Website 1 (N = 468)	Website 2 (N = 185)	Website 3 (N = 100)	Website 4 (N = 607)
Ease of use	3.60 (.75)	3.56 (.81)	3.57 (.83)	3.50 (.83)
Hyperlinks	3.39 (.68)	3.40 (.77)	3.39 (.73)	3.26 (.76)*
Structure	3.46 (.64)	3.47 (.71)	3.44 (.69)	3.35 (.72)
Relevance	3.87 (.56)	3.87 (.62)	3.87 (.63)	3.87 (.55)
Comprehension	3.89 (.49)	3.92 (.52)	3.98 (.55)	3.90 (.49)
Completeness	3.54 (.60)	3.53 (.69)	3.50 (.77)	3.49 (.64)
Lay out	3.52 (.75)	3.58 (.76)	3.56 (.78)	3.48 (.79)

Note: Opinions were measured on a five-point scale, where 1 is most negative and 5 is most positive (negatively stated items were reversed).

4. General conclusion and discussion

Questionnaires that measure the quality of governmental websites are frequently used in practice, but are not often based on sound research. In this article we advocated a sound evaluation of website questionnaires on six issues: (1) general availability of the questionnaire, (2) clarity concerning the domain the questionnaire can be applied to, (3) the goals of the questionnaire, (4) the underlying factor structure, (5) the stability and reliability of this structure, and (6) the sensitivity to differentiate between websites.

The Website Evaluation Questionnaire, a multidimensional instrument for assessing the quality of governmental websites, proved to be a valid and reliable questionnaire. The WEQ has seven clearly distinct dimensions which measure website quality in a stable way over different websites. In a controlled laboratory setting the questionnaire is sensitive to differences between websites. However, some remarks must be made for online evaluation with the WEQ. The online survey partly measures something different than in the controlled setting, which is indicated by the higher correlations between the dimensions. In the online setting the WEQ is less sensitive to differences between websites than in the controlled setting. This may firstly be explained by the broadness of the municipal websites we evaluated in our studies. In the laboratory condition we limited the evaluations to two or three scenarios which were carried out by the users, who thus all based their judgments on the same parts of the websites. In the online condition there was no control over the pages users visited and the goals they had. This means that the object that was judged differed more in the online setting than in the controlled setting. The diverse experiences online users based their opinions on, made it difficult to reveal differences between the websites. A second explanation may be the more global way users filled out the questionnaire online. Users seem to have based their judgments on a general impression of the website, which means that they were influenced by categories other than the specific category that had to be evaluated, the so called halo-effect.

The results of study 2 show that with an online measurement of extensive governmental websites, only large differences between websites may be revealed because of a tendency to the mean. It seems difficult to distinguish the more subtle diversity in website quality. This finding not only applies to the questionnaire we tested, but also is part of a more fundamental issue. It is debatable whether it is sensible to measure one diffuse entity with a questionnaire and thus generalize over everything on the website. However, governmental organizations need overall indicators for website quality, and the questionnaire is very suitable for a quantitative presentation of user opinions. We should preserve the valuable insights that a questionnaire can provide and at the same time measure website quality more precisely. This can be done by relating the results of the questionnaire to the actual behavior of website users. The most accurate way to do this is by including server logs in the analysis, which will give more insight into the experiences and pages users base their opinions on. A more easily applicable solution is to place the questionnaire on a selection of 'end pages' on the website, which makes it possible to relate opinions to certain goals and pages. In this way the presumption that a questionnaire measures the quality of the entire website should be relinquished and quality judgments should only be related to the parts of the website that were actually visited. Another solution is to add open ended questions to the questionnaire concerning users' goals and the pages they visited. A drawback of this solution is that the analysis of the data will be more complicated and that some parts of the website will get too few respondents to draw adequate conclusions.

This research shows that a sound questionnaire that is able to discriminate between websites in laboratory settings, may have difficulties detecting differences online. One conclusion we may draw from this finding is that differences measured online will be very meaningful. In our study we did not select the websites on quality differences beforehand, which means that the four municipal websites we tested may not differ from each other substantially. A follow up study (to be published) with eight other online municipal websites indeed shows that the four websites in our study hardly differ from each other, while the WEQ does reveal several significant differences between the other eight websites and also between different versions of websites.

The reliability of the questionnaire remains an issue that deserves attention. In most of the studies described in this paper, the *comprehension* dimension had a reliability estimate lower than .70. In one of the evaluations and in earlier studies, this dimension, with the same three items, received higher estimates. Reliability estimates depend on the sample of respondents and on the proportion of true score variance, which means that the estimates are not always stable over measurements. This makes it important to monitor the reliability constantly, and to not be too easily satisfied when a reliability estimate is high in one study.

The representativeness of the sample is an important issue in all evaluations, especially in an online setting. However, the characteristics of the target population are not always evident. In our laboratory study, we chose for an equal distribution of the respondents over three educational levels. In the online study, however, the sample consisted of 57% respondents with a higher educational level, and only 9% lower educated respondents, which might be a limitation of this study. However, it is doubtable which sample is more representative. We know that higher educated people visit governmental websites more frequently than lower educated people (see also Section 3.1.1), but no information is available on the exact characteristics of the websites' visitors. The online sample might therefore be more adequate than the laboratory sample.

In this study we focused on the measurement of government website quality from a usability point of view. However, user opinions on the usability of these websites might be influenced by other factors that we did not measure, such as the extent to which citizens trust their government, political factors, or expectations based on earlier experiences on governmental websites. Research has shown that trust is an important factor for people's adoption of electronic services (Akkaya, Wolf, & Krcmar, 2010; Beldad, De Jong, & Steehouder, 2010), and it can be expected that trust also influences users' opinions on the usability. Also, users' expectations seem to influence their usability ratings (Raita & Oulasvirta, 2011). More research is needed to gain insight into the factors that might influence users' opinions on usability in a governmental context.

As stated earlier, it is very important to evaluate governmental websites and to further improve their quality. This study has thoroughly tested one instrument that can be used for such evaluations: the Website Evaluation Questionnaire proved to be a valid and reliable instrument. Future research should further exploit the strengths and weaknesses of different kinds of evaluation methods, ranging from specific to more generic, expert-focused to user-focused, and qualitative to quantitative. Also, more research is needed on combinations of different types of methods, which can complement each other and in this way contribute to a higher standard of governmental website evaluation.

Acknowledgments

This article is based on a research project financed by the Dutch Organization for Scientific Research (NWO). It is part of the research program Evaluation of Municipal Web Sites. We thank the anonymous reviewers for their valuable comments on earlier versions of this article.

Appendix A. The Website Evaluation Questionnaire (WEQ)

The questions are presented by dimension. The sequence of the questions used in evaluations, is showed with the numbers behind the questions (between brackets).

1. Ease of use
1. I find this website easy to use (4)
2. I had difficulty using this website (11)
3. I consider this website user friendly (18)

2. Hyperlinks
4. The homepage clearly directs me towards the information I need (5)
5. The homepage immediately points me to the information I need (12)
6. It is unclear which hyperlink will lead to the information I am looking for (19)
7. Under the hyperlinks, I found the information I expected to find there (22)

3. Structure
8. I know where to find the information I need on this website (6)
9. I was constantly being redirected on this website while I was looking for information (13)
10. I find the structure of this website clear (20)
11. The convenient set-up of the website helps me find the information I am looking for (23)

4. Relevance
12. I find the information in this website helpful (1)
13. The information in this website is of little use to me (8)
14. This website offers information that I find useful (15)

5. Comprehension
15. The language used in this website is clear to me (2)
16. I find the information in this website easy to understand (9)
17. I find many words in this website difficult to understand (16)

6. Completeness
18. This website provides me with sufficient information (3)
19. I find the information in this website incomplete (10)
20. I find the information in this website precise (17)

7. Lay out
21. I think this website looks unattractive (7)
22. I like the way this website looks (14)
23. I find the design of this website appealing (21)

8. Search option[a]
24. The search option on this website helps me to find the right information quickly (24)
25. The search option on this website gives me useful results (25)
26. The search option on this website gives me too many irrelevant results (26)

[a] The WEQ may be complemented with questions on the search option. These questions were not relevant in the laboratory study and were therefore not included in the analyses in this article.

References

Akkaya, C., Wolf, P., & Krcmar, H. (2010). The role of trust in e-government adoptation: A literature review. *AMCIS 2010 proceedings, paper 297*.

Albert, W., & Dixon, E. (2003). Is this what you expected? The use of expectation measures in usability testing. *Proceedings of the usability professionals association 2003 conference* Scottsdale, AZ.

Anderson, E. W., & Fornell, C. (2000). Foundations of the American Customer Satisfaction Index. *Total Quality Management*, *11*, 869–882.

Bangor, A., Kortum, P., & Miller, J. A. (2008). The System Usability Scale (SUS): An empirical evaluation. *International Journal of Human Computer Interaction*, *24*, 574–594.

Bangor, A., Kortum, P., & Miller, J. A. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123.

Beldad, A., De Jong, M., & Steehouder, M. (2010). How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust. *Computers in Human Behavior*, *26*, 857–869.

Bertot, J. C., & Jaeger, P. T. (2008). The E-Government paradox: Better customer service doesn't necessarily cost less. *Government Information Quarterly*, *25*, 149–154.

Brinkman, W. P., Haakma, R., & Bouwhuis, D. G. (2009). The theoretical foundation and validity of a component-based usability questionnaire. *Behaviour and Information Technology*, *28*, 121–137.

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry*. London: Taylor and Francis, http://www.usability.serco.com/trump/documents/Suschapt.doc

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 464–494.

Couper, M. P., & Miller, P. V. (2008). Web survey methods: Introduction. *Public Opinion Quarterly*, *72*, 831–835.

De Jong, M., & Lentz, L. (2006). Scenario evaluation of municipal websites. Development and use of an expert-focused evaluation tool. *Government Information Quarterly*, *23*, 191–206.

Elling, S., Lentz, L., & De Jong, M. (2007). Website Evaluation Questionnaire: Development of a research-based tool for evaluating informational websites. *Lecture Notes in Computer Science*, *4656*, 293–304.

Elling, S., Lentz, L., & De Jong, M. (2011). Retrospective think-aloud method: Using eye movements as an extra cue for participants' verbalizations. *Proceedings of the 2011 annual conference on human factors in computing systems* (pp. 1161–1170). New York: ACM.

Elling, S., Lentz, L., & De Jong, M. (2012). Users' abilities to review website pages. *Journal of Business and Technical Communication*, *26*, 170–200.

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, *26*, 132–139.

Galesic, M., & Bosjnak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*, 349–360.

Heerwegh, D., & Loosveldt, G. (2008). Face-to-face web surveying in a high-internet-coverage population. Differences in response quality. *Public Opinion Quarterly*, *72*, 836–846.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human Computer Studies*, *64*, 79–102.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109–132.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8.51*. Chicago: Scientific Software International.

Kirakowski, J., Claridge, N., & Whitehand, R. (1998). Human-centered measures of success in website design. *Proceedings of the 4th conference on human factors & the web* (pp. 1–9). Baskerville, NJ.

Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *SIGGHI Bulletin*, *23*(1), 78–81.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the System Usability Scale. *Lecture Notes in Computer Science*, *5619*, 94–103.

Loukis, E., Xenakis, A., & Charalabidis, Y. (2010). An evaluation framework for e-participation in parliaments. *International Journal of Electronic Governance*, *3*, 25–47.

Magoutas, B., Halaris, C., & Mentzas, G. (2007). An ontology for the multi-perspective evaluation of quality in e-government services. *Lecture Notes in Computer Science*, *4656*, 318–329.

McGee, M. (2004). Master usability scaling: Magnitude estimation and master scaling applied to usability measurement. *Proceedings of CHI conference on human factors in computer systems* (pp. 335–342). Vienna, NY: ACM.

Muylle, S., Moenaert, R., & Despontin, M. (2004). The conceptualization and empirical validation of website user satisfaction. *Information Management*, *41*, 543–560.

Ozok, A. A. (2008). Survey design and implementation in HCI. In J. Jacko, & A. Sears (Eds.), *Handbook of human–computer interaction* (pp. 151–1169). New York: Lawrence Erlbaum.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*, 74–92.

Raita, E., & Oulasvirta, A. (2011). Too good to be bad: Favorable product expectations boost subjective usability ratings. *Interacting with Computers*, *23*, 363–371.

Rijkeboer, M. M., & Van den Bergh, H. (2006). Multiple group confirmatory factor analysis of the Young Schema-Questionnaire in a Dutch clinical versus non-clinical population. *Cognitive Therapy and Research*, *30*, 263–278.

Sauro, J., & Dumas, J. (2009). Comparison of three one-question, post-task usability questionnaires. *Proceedings of the 27th international conference on human factors in computing systems* (pp. 1599–1608). Boston.

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., Wei, C., Barrick, J., Cuddihy, E., & Maust, B. (2005). Internet-based research: Providing a foundation for web-design guidelines. *IEEE Transactions on Professional Communication*, *48*, 242–260.

Tedesco, D., & Tullis, T. (2006). A comparison of methods for eliciting post-task subjective ratings in usability testing. *Usability Professionals Association (UPA) conference* (pp. 1–9). Denver, Colorado.

Tourangeau, R. (2003). Cognitive aspects of survey measurement and mismeasurement. *International Journal of Public Opinion Research*, *15*, 3–7.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Tullis, T., & Albert, B. (2008). *Measuring the user experience. Collecting, analyzing, and presenting usability metrics*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Tullis, S. T., & Stetson, J. N. (2004). A comparison of questionnaires for assessing website usability. *Usability Professionals Association (UPA) Conference* (pp. 7–11). Minneapolis, USA.

Van den Haak, M. J., De Jong, M. D. T., & Schellens, P. J. (2007). Evaluation of an informational website: Three variants of the think-aloud method compared. *Technical Communication*, *54*, 58–71.
Van den Haak, M., De Jong, M., & Schellens, P. J. (2009). Evaluating municipal websites: A methodological comparison of three think-aloud variants. *Government Information Quarterly*, *26*, 193–202.
Van Deursen, A. J. A. M., & Van Dijk, J. A. G. M. (2009). Improving digital skills for the use of online public information and services. *Government Information Quarterly*, *26*, 333–340.
Van Dijk, J., Pieterson, W., Van Deursen, A., & Ebbers, W. (2007). E-services for citizens: The Dutch usage case. *Lecture Notes in Computer Science*, *4656*, 155–166.
Van Schaik, P., & Ling, J. (2005). Five psychometric scales for online measurement of the quality of human–computer interaction in websites. *International Journal of Human Computer Interaction*, *18*, 309–322.
Verdegem, P., & Verleye, G. (2009). User-centered e-government in practice: A comprehensive model for measuring user satisfaction. *Government Information Quarterly*, *26*, 487–497.
Welle Donker-Kuijer, M., De Jong, M., & Lentz, L. (2010). Usable guidelines for usable websites? An analysis of five e-government heuristics. *Government Information Quarterly*, *27*, 254–263.
Young, J. E. (1994). *Cognitive therapy for personality disorders. A schema-focused approach (revised ed.).* Sarasota, FL: Professional Resource Press.
Zijlstra, F. (1993). *Efficiency in work behavior. A design approach for modern tools.* PhD Thesis, Delft University of Technology. Delft, The Netherlands: Delft University Press.

Sanne Elling is a PhD-student at the Utrecht Institute for Linguistics UIL-OTS at Utrecht University in The Netherlands. Her research project is on user focused methods of website usability evaluation.

Leo Lentz is a Professor at the Utrecht Institute for Linguistics UIL-OTS at Utrecht University in The Netherlands. Web Usability and Text evaluation are the main focus of his research.

Menno de Jong is a Professor of Communication Studies at the University of Twente in The Netherlands. His main research interests include the methodology of applied research techniques. He has published research articles on various methods of usability evaluation.

Huub van den Bergh is a Professor of teaching and testing of language proficiency at Utrecht University. His research is on cognitive mechanisms of text interpretation and text production. The focus of his research is on writing research and methodological features of measuring processes of text interpretation and production.