Master's thesis
Faculty for Psychology
University of Basel

Handed in October 15, 2011

FUS - Form Usability Scale

Development of a Usability Measuring Tool for Online Forms

Thesis Supervisors:

Orsini Sebastien, M. Sc.

Department of Cognitive Psychology and Methodology, University of Basel

Prof. Dr. Klaus Opwis

Department of Cognitive Psychology and Methodology, University of Basel

Andreas Aeberhard
Matr. Nr.: 05-058-763
Rufacherstr. 30
CH-4055 Basel
+41 79 516 52 82

e-mail: a.aeberhard@unibas.ch

Abstract

In recent years, the use of online forms has become increasingly important and more common on the internet. More and more companies are using online forms to obtain customer information. In this context the question of the form's quality arises: How can we measure user satisfaction of online forms? In this work a questionnaire is presented that measures user satisfaction and the usability of online forms. Developing steps are shown. The questionnaire, containing ten questions, was tested ($N = 92$) in a laboratory setting with six forms. A high internal consistency (Cronbach α) of .85, moderate item difficulties (.51 to .93) and good discriminatory power coefficients (.145 to .861) and a good homogeneity range (.237 to .578) was found. One of the ten questions did not meet the statistical requirements, resulting in a final set of nine questions.

Contents

Introduction

Forms are everywhere nowadays (Jarrett & Gaffney, 2008). We face them on a daily basis so much so that some of them have become part of the ”norm”; we do not perceive them as forms at all anymore. One example of this is the login form on a computer system, where a username and password is entered. These are unlike forms which may never be mistaken, such as marriage certificates and wills.

More and more companies offer their products and services online for sale – worldwide (Brynjolfsson & Smith, 2003). Due to the growth in e-commerce, there has been a rise in the use of online forms by companies to acquire personal information from their customers. This necessity will lead to a greater amount of online forms, day by day and according to Niedermann and Uhr (2008) a usable online form leads to more customers, increasing the company's income.

With this development in mind, the quality and usability of online forms will become evermore crucial and important. This requires a construct which gives the possibility to check the quality and usability in a fast and reliable way. The main problem that is to be solved is the reluctance of every user to fill in forms (Jarrett & Gaffney, 2008). Therefore the question is: How do we reduce this reluctance and thus maximize the companies' profit? The answer is: With a form that is as user-friendly as possible.

Tullis and Stetson (2004) showed that this trend was recognized early and measuring tools to evaluate the quality and usability have been in the process of development for years. This was, of course, not without good reason: the consequences of a bad form could be critical. In a study conducted by Hoffmann, Zimmerman, and Tompkins (1996), 41% of all persons gave contradictory instruction in their living wills. These contradictory instructions were caused by the bad form of the living will which they had to fill in.

Good forms, on the other hand, may have desirable effects; evidence for this is shown by the redesign of eBay's registration forms in 2002. In March 2002 there were 46 million registered users on eBay this increased greatly after a new, more user-friendly, registration form was created in late 2003; the number of registered users grew to 95 million. This is an increase of more than 100% within a two years which led to an increase in sales too (Herman, 2004).

These two examples show that it is crucial to focus on the usability aspect of a form from both a social and an economic point of view. Nielsen (2002) stated that the user is less willing to put up with a difficult and complicated interface because he interacted with good interfaces once and knows that it is possible for an interface to be good. Further, Nielsen (2002) has concluded that high usability is desirable. For example, a user encounters a unusable order form in an online bookstore. He will probably discard his order and go to an other online bookstore with a more usable order form than the one just encountered.

The aim of this work was to develop a *Form Usability Scale* (FUS), through a structured questionnaire that measures the usability of online forms. To achieve this goal, we tested the FUS in a laboratory setting to gather both objective and subjective data. Conclusions towards validity and reliability were drawn based on the data collected; in addition an item analysis was conducted based on classic test construction theories.

Theoretical Background

An ”online form” is defined as ”a web page that has boxes you can type into.” (Jarrett & Gaffney, 2008). Additionally, a form may have radio-buttons, drop-downs and checkboxes (Jarrett & Gaffney, 2008). Data forms are commonplace on the internet, popular examples include: the registration forms for online shops and message boards, age verification on websites, the login for webmail and even Google's searchbox. This work assumes that users on the internet have a certain knowledge and experience in using online forms and that they know how to handle them properly.

The range in quality of these online forms is huge, it reaches from forms that are very usable to those which are completely inaccessible to the user; this is because there are currently no universally recognized guidelines that provide a framework for programmers and designers to design online forms. Nevertheless, in the last decade some recommendations were published from several researchers and authors (for an overview see Koyani (2006), Spool (1999) and James, Beaumont, Stephens, and Ullman (2002)). Most of these recommendations should, in principle, increase the online form's usability and thus increase user satisfaction. It is assumed that users are able to distinguish between good and bad online forms from the usability point of view; this is supported by the previously made assumption about the knowledge and experience of the user.

Each user reacts differently to the online forms that they encounter. The technology aversive retiree shows a different reaction if he encounters an online form at an online shop than a technology affine computer science student (Jarrett & Gaffney, 2008). Now, if one alters the usability of these online forms, the reactions of the retiree and the computer science student will differ a lot more. It is probable that the retiree will lose interest and leave the online form if it is unusable to him, whereas the computer science student is likely to have more tolerance and sufficient expertise to persevere and complete an unusable online form.

That is why attention should be drawn towards a broad distribution of user's demographics so that there is a balance between technology aversive and technology affine users.

Based on their different behaviours, Jarrett and Gaffney (2008) grouped users into three roles: (1) Readers, (2) Rushers and (3) Refusers. The Readers read the forms carefully, the

Rushers fill in the form without reading it and only read it if they think it is inevitable and the Refusers do not want to interact with the form in any way and will leave it. In our case, the retiree is a Reader and the computer science student a Rusher. These roles vary between different kind of forms (Jarrett & Gaffney, 2008), for example the student will become a Reader if he fills out the tax declaration.

A company or an institution aims at a low quote of Refusers and thus maximize the number of potential customers. To acquire this goal they should stick to the following three rules with their online forms: (1) Establish trust, (2) Reduce social cost and (3) Increase reward. A high user satisfaction has a direct influence to the first and second rule and will result in fewer Refusers (Jarrett & Gaffney, 2008).

We assume that the FUS could be able to identify these roles. Readers and Rushers can be identified by objective data (time needed to fill in the form) and subjective data (ratings). Refusers are harder to identify because they do not want to interact with the online form and leave it in the first place. Nevertheless, it should be possible to identify future Refusers by evaluating subjective data and comments of a user. A user that is very unhappy with the form provides bad ratings and is likely to provide a comment to a question.

The definition of ”user satisfaction” is contextual; when working with a computer, user satisfaction means that the user is able to work prolifically. However, productivity plays a lesser role in user satisfaction when playing a computer game or surfing the web. The International Organization for Standardization (ISO) categorizes user satisfaction with efficiency and effectiveness as the three pillars of system usability (DIN, 1998).

According to Hassenzahl, Beu, and Burmester (2001), the main research focus of the human-computer interaction (HCI) literature is on efficiency and effectiveness and user satisfaction is seen as a by-product of good usability, although user satisfaction is a key component in many situations and not only a by-product (Lewis, 1995).

Usability plays an important role in HCI (Hornbæk, 2006), and is the most popular construct used to predict and measure the success of an information system (Huang, Yang, Jin, & Chiu, 2004). Many different researchers have offered definitions (see Bevan (1995) and

Shackel (1991) for an overview) whereby most agree that usability is contextual (Newman & Taylor, 1999). Quesenbery (2001) defines usability through four key principles: ”(1) Usability means thinking about how and why people use a product, (2) Usability means evaluation, (3) Usability means more than just "ease of use" and (4) Usability means user-centered design.”

The last key principle implies that the user will be satisfied if their: (1) goals, (2) mental model, (3) tasks and (4) requirements are all met. At the same time, a product is usable if (1) analysis, (2) design and (3) evaluation are conducted from the perspective and point of view of the user, thereby ensuring that all four points have been fulfilled (Quesenbery, 2001).

There are several questionnaires that allow a programmer to assess the user satisfaction of an online form prior to it’s public launch. Four popular examples include the (1) *Computer System Usability Questionnaire* (CSUQ, Lewis (1995)), the (2) *Intranet Satisfaction Questionnaire* (ISQ, Bargas-Avila, Lötscher, Orsini, and Opwis (2009)), the (3) *System Usability Scale* (SUS, Brooke (1996)) and the (4) *Questionnaire for User Interface Satisfaction* (QUIS, Chin, Diehl, and Norman (1988)).

The CSUQ, tested and validated with 377 subjects (Lewis, 1995), was developed by IBM and consists of nineteen questions (Lewis, 1995) which are used as a basis by to evaluate the usability of a computer system. These same nineteen questions from the PSSUQ (Lewis, 1991) are used for the CSUQ, with a single wording difference (Lewis, 1995). Initially, the CSUQ was developed for computer systems, however it was later adapted and re-phrased for use on websites (Tullis & Stetson, 2004). The questions from the CSUQ can be answered on a seven-point Likert-Scale with preferential options ranging from ”Strongly Disagree” to ”Strongly Agree”. Furthermore, this scale forces the user to answer the question as there is no ”I cannot anwer this question” option available to respond with.

The ISQ was developed by Bargas-Avila et al. (2009) and consists of eighteen questions which are used to evaluate the user satisfaction of employees utilizing a company’s intranet. Like the CSUQ, the ISQ uses a Likert-Scale to assess user ratings; however unlike the CSUQ, it uses a six-point Likert-Scale. In the event that the user does not know how to answer one of the questions an additional option, ”I cannot answer this question”, was made available.

The ISQ was tested and validated twice with 881 subjects the first time and 1350 subjects the second (Bargas-Avila et al., 2009).

The SUS was developed by Brooke (1996) and consists of ten questions which are used by to evaluates the global view of subjective assessments of a system's usability. It uses a five-point Likert-Scale to assess user ratings, ranging from "Strongly Disagree" to "Strongly Agree". As with the CSUQ, the user is made to answer the question because there is no option to not answer like that offered by the ISQ. The SUS is a robust and reliable model that correlates well with other questionnaires that assess usability on a subjective basis (Brooke, 1996). The SUS was tested and validated with 2,324 surveys (Bangor, Kortum, & Miller, 2008).

The QUIS was developed by Chin et al. (1988) and consists of twenty-seven questions which allow the QUIS to measure the user's subjective rating of an interface. A ten-point Likert-Scale is used to assess a user's approval an interface. Dissimilarly to the other questionnaires, the QUIS does not rely on a "Strongly Disagree" to "Strongly Agree" scale, and instead uses descriptive phrases that range from "helpful" to "unhelpful" and "always" to "never". Again, the user has to answer the question. The QUIS was tested and validated with 4,597 subjects (Chin et al., 1988).

User satisfaction can be measured using the FUS, and therefore allows statements to be made in regard of success resp. failure of the online form. However, operationalisation of user satisfaction, is no simple task; this is made especially difficult as there is no agreement between researchers on *how* user satisfaction can be operationalised (DeLone & McLean, 1992; Goodhue, 1993; Hamilton & Chervany, 1981; Ives & Olson, 1984; Miller & Doyle, 1987; Shirani, Aiken, & Reithel, 1994; Symons, 1991). Additionally, in many studies there is no theoretical background on how this process of operationalisation has been done (Melone, 1990).

The current work defines user satisfaction as follows: "user satisfaction is the subjective sum of the interactive experience." (Lindgaard & Dudek, 2003). This means that the user will be satisfied if the interaction experience amounts. Usability has an influence on the success of information systems (Huang et al., 2004) and, due to the fact that user satisfaction is one of the three pillars of the usability, it also has influence on the success of that system (Al-Khaldi & Wallace, 1999; Szajna & Scamell, 1993). Thus, it is the goal of every company to design an online form that is as usable as possible, highlighting the necessity for a reliable and valid questionnaire for online forms.

The usability of an interface plays a major role in user satisfaction; low usability has a negative impact on user satisfaction whereas high usability has a positive impact (Park & Hwan Lim, 1999; Frøkjær, Hertzum, & Hornbæk, 2000). Therefore, the user will be more satisfied if the online form that they have filled in is deemed usable than the user that experienced an unusable online form.

In regard to the FUS, it is assumed that if a form has a high user-rating then the four user satisfaction key requirements have been fulfilled and that the user is satisfied. On the other hand, if a form has a low rating, the four points have, assumedly, *not* been fulfilled and the user remains unsatisfied. Additionally, it is easier to implement the three points from the user's point of view because it is possible on the basis of the ratings to see which aspect of the online form are problematic for the user.

Considering the nature of online forms, one must remain aware of guidelines mentioned in the introduction. The study of Bargas-Avila et al. (2010) summarizes twenty guidelines for a usable online form. If these twenty guidelines are followed, the online form should become more usable. Questions in the FUS are inspired by these guidelines.

The FUS combines the aforementioned aspects and measures usability and user satisfaction effectively in a short space time. Most of the current questionnaires focus on just one of these two constructs. Additionally, it can be used in an online environment and therefore is handy for companies to assess their online forms on a grand scale with little effort. In doing so, they are able to design more usable and satisfying online forms, which theoretically leads to an increase in potential customers and therefore, more sales. Most of the current questionnaires use around twenty questions to conduct their construct, leading to fewer voluntary participants in the questionnaire.

At this time, these observations lead to the conclusion that there is no questionnaire that fulfills all of these aspects and that there is a need for a new questionnaire that does. The FUS aims to fulfill these aspects and fill the gap.

Method

The current questionnaire emerged through several steps. The first point of action was to create questions based on the aforementioned theoretical background. Some of the questions have been modified or reduced due to various reasons. In the following section these steps will be explained and described. Furthermore, there will be a closer look at methodological aspects and test arrangements.

Construction of the Questionnaire

Prior to the construction of this questionnaire other usability questionnaires were examined; based on the aforementioned questionnaires, the first draft of the FUS was created.

The CSUQ and ISQ along with current research literature, influenced the first draft of the FUS. Additionally, other questionnaires like the SUS and the QUIS had a small influence too. Also, impact was created by the study of Bargas-Avila et al. (2010) with it's 20 guidelines. All of these questionnaires and the guidelines influenced the creation of the FUS, resulting in a first draft containing 20 questions.

The first draft was sent to five experts to be reviewed. The questionnaire was modified and some questions were deleted based on their feedback, this lead to the creation of the actual questionnaire. The main focus of the feedback was concerning the wording of the questions and their similarity. The questions were reduced and the wording was adapted. After several alterations, the final version containing ten questions emerged. These ten questions were chosen with the specific intent to cover of all aspects of a form. Table 1 shows these ten questions. Note that the testing, later described, was conducted using these ten questions.

Mind that special attention was given towards the number of questions used in the FUS. Batinic (2000) stated that a user is willing to spend 6 to 15 minutes to answer an online questionnaire. The FUS, consisting of ten questions, will take 5 to 8 minutes to be filled out and therefore fulfills the recommendations by Batinic (2000).

Table 1: The Ten Questions of the FUS, Final Version. Translated by the Author.

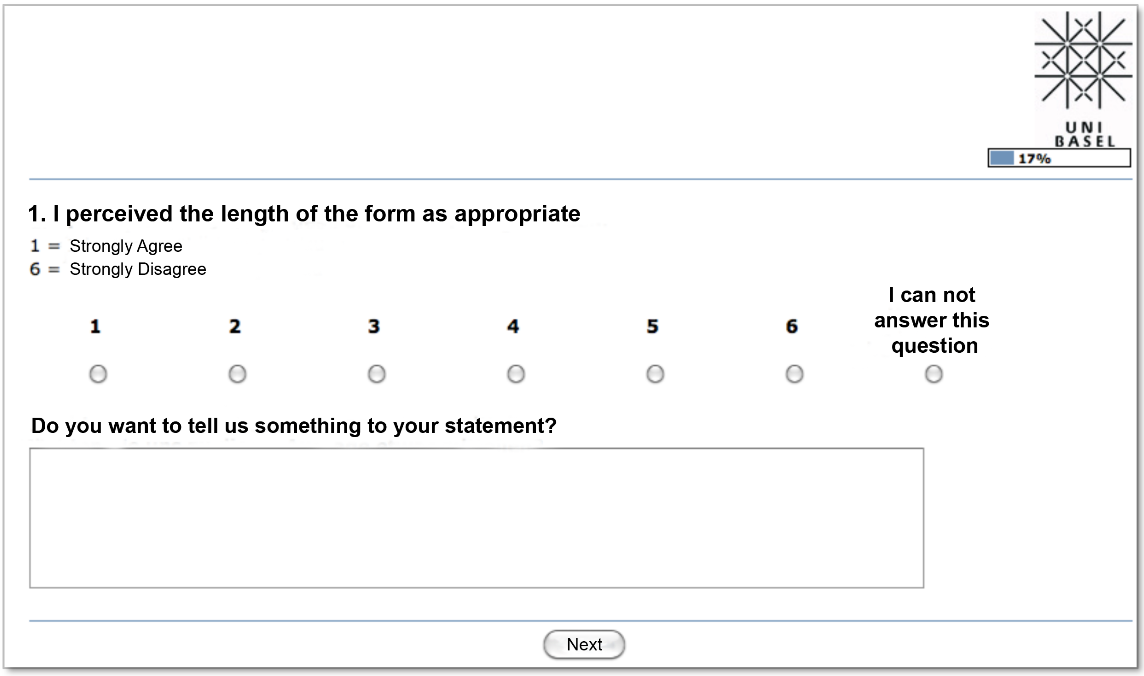
1. I perceived the length of the form as appropriate.
2. I was able to fill in the form quickly.
3. I perceived the order of the questions in the form as logical.
4. Mandatory fields were clearly visible in the form.
5. I knew all the time which information were expected from me.
6. I knew at every input what rule I had to stick to (e.g. possible answer length, date format or password requirements).
7. The fill in was eased by given answers (e.g. drop-down menus, checkboxes etc.)
8. In case of a problem I was instructed by a error message how to solve the problem. (Please check "I can not answer this question" if there were no problems)
9. Purpose and utility of the form was clear.
10. In general I am pleased with the form.

Scale

To answer the questions, a six-point Likert-Scale has been used with the additional option, "I cannot answer this question". The scale reaches from 1 = "Strongly Disagree" to 6 = "Strongly Agree". The scale was continuously numbered whereas only number 1 and 6 were labelled to ensure that it was an interval scale. See figure 1 for a generic screenshot of the questions.

According to Borg (2002), the advantage of a six-point Likert-Scale is a higher reliability and validity compared to a five or seven-point Likert-Scale, it was to this basis that a six-point Likert-Scale was used for the FUS. Additionally, an even number of possible answers was chosen with the effect that no neutral answers could be given. Mummendey (1995) asserts that there are several reasons why users choose neutral answers: (1) they really have a neutral opinion, (2) they do not know how to answer the question, (3) they think the question is irrelevant, (4) they refuse to answer the question or (5) they want to express their reluctance towards the question. A motivated user will avoid the neutral answer (Rost, 2004).

The subjects were able to voice their opinion with every question by filling out the comment field below the Likert-Scale. These comments can be used for qualitative statements or for interpreting data. A progress bar in the top right corner enabled the subjects to see how close they were to completing the FUS (see figure 1). Subjects also had the option to leave the form blank and continue to the next question as there was no input control.



1. I perceived the length of the form as appropriate

1 = Strongly Agree
6 = Strongly Disagree

1	2	3	4	5	6	I can not answer this question
○	○	○	○	○	○	○

Do you want to tell us something to your statement?

Figure 1. Screenshot of the Survey. Here Question 1 is an Example. Translated by the Author.

Testing

The Canton of Basel-Stadt agreed to co-operate and provide it's online forms for testing. There were two kinds of forms provided by the Canton of Basel-Stadt: three online forms that are currently in use (henceforth named "old forms") and three online forms that are in development (henceforth named "new forms"). All six forms are used by the population to help change personal information. Thus, a person does not fill in these kinds of forms very often. The service in return was a usability testing of the old and new forms, including all gathered raw data and improvement suggestions for the new forms.

A usability test was conducted in a laboratory setting in the hope of gathering objective data by using the eye tracker. This objective data was (1) the time taken and necessitated by each subject to fill a form out, (2) the amount of fixations on the screen and (3) the amount of mouse clicks taken to complete the form. This objective data was used and compared with subjective data to evaluate the FUS.

Acquiring the Sample Survey. All subjects received an email with an invitation to participate in the test. Overall, 600 people contacted randomly and all of the subjects came from the database of the Faculty for Psychology, Basel, where they had registered to receive study invitations. Subjects were able to choose a testing date by selecting a date when they would be available, entering their name, email address and mobile phone number. The scheduling was organized on a first come, first served basis.

Elicitation and Testsetting. The testing was conducted over a period of two months. In the laboratory the subjects had been orally introduced to the study and were given a brief instruction about the eye tracker. The subjects were encouraged and asked to think aloud during the whole testing process.

Additionally, subjects received a hard-copy manual which included an overview of the whole project and its procedure. Moreover it contained questions about demographics and computer skills and all of the information needed to fill the forms in (fake address, birthdate, email address, situations, telephone number and job title).

Once the subject filled out their demographics in the manual, the first form was presented. After each form the subject answered the ten FUS questions, therefore completing the FUS a total of six times.

There were two types of testing; the first started with a new form followed by an old one whereas the second version started with an old form followed by a new one. Old and new forms were tested in an alternative order to avoid the possibility of sequence effects. See figure 2 for a schematic figure.

Participants

Overall 92 subjects participated in the study. Three-quarters (68 = 73.9%) were female and one quarter (24 = 26.1%) were male. The average age was 29.3 years with a standard deviation of 11.73. The age ranges from 16 to 63 years. Figure 3 visualizes the age distribution.

Practically everyone (91 = 98.9%) uses the internet at least once a day and so it can be assumed that every subject had filled out an online form at least once (e.g. to subscribe to

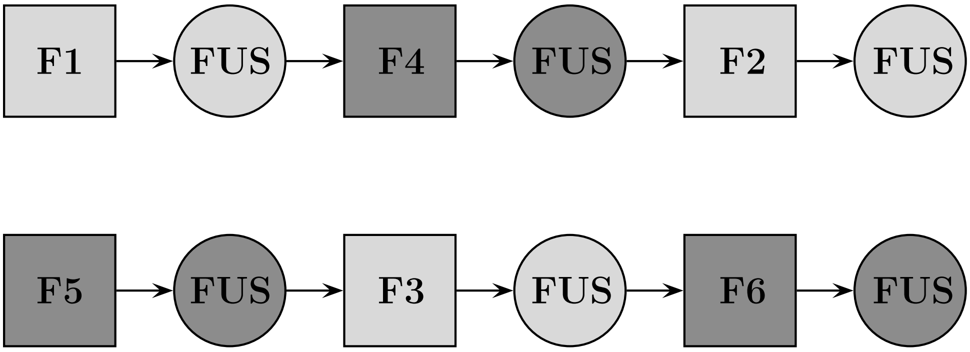


Figure 2. Schematic Figure of the Testsetting. New Forms are Darkgray, Old Forms are Gray. F1 to F6 are the Different Tested Forms.

the subject database which they were recruited from). Only half of the subjects visited the website of the Canton Basel-Stadt before the testing. Overall 6 (= 6.5%) subjects explicitly stated that they had filled in forms on the website of the the Canton Basel-Stadt.

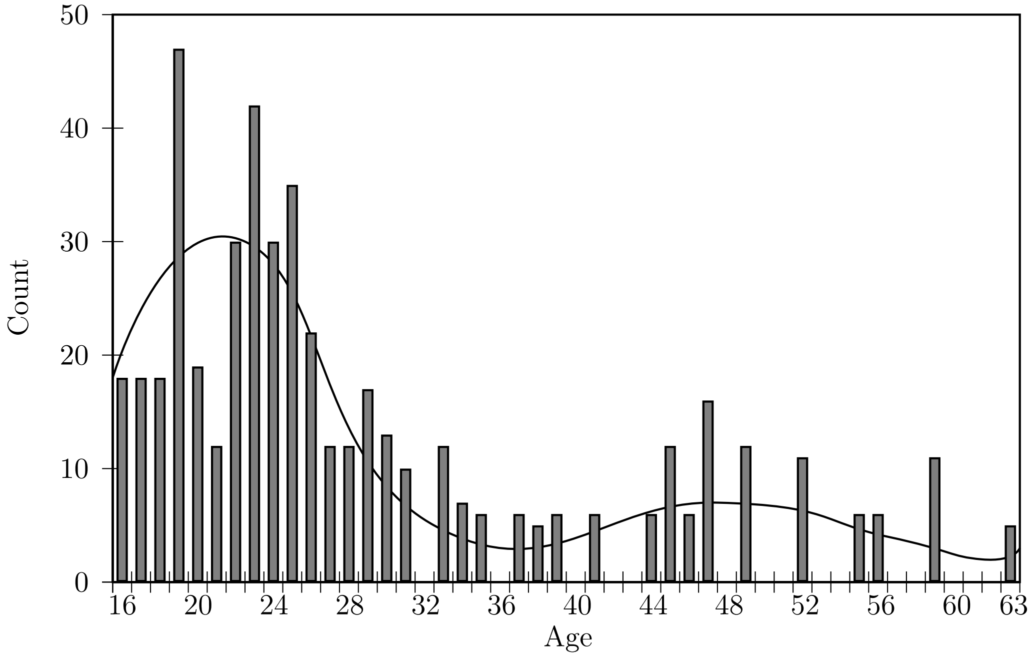


Figure 3. Age Distribution Among All Subjects, Mean=29.3, SD=11.73, Range=16–63 Y.

Results

Initially, all dropouts were excluded. In every data row answers with a 6, a 1 or a missing (no answer) were added together to a overall data row score which lies between 0 and 10. Data rows with a score 9 or 10 were deleted because (1) the subject did not differ in his answer or (2) too many answers were not answered or (3) both.

With an overall score of 8 or less it is possible to argue that the subject differs in answering the questions. By this reasoning, the threshold was set to the overall score of 8 leading to an exclusion range of data rows of 8.6% in form 1 (the most) to 2.2% in form 3 (the least). Overall 5.5% of all data rows were excluded using the described method. Table 2 shows the exclusion count per form.

Missing Value. The answer ”I can not answer this question” was also included as a missing value. The missing value quota was below 2.0% for seven out of the ten questions. For three questions the missing quota, 5%, was exceeded considerably. The highest missing quota lies in question 8 with 80.4% (see table 3). This exceedingly high missing quota is,

Table 2: Exclusions per Form

	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6	Total
N	93	90	90	92	92	89	546
N Excluded	8	6	2	3	6	5	30
% Overall	8.6	6.7	2.2	3.3	6.6	5.6	5.5
N after Exclusion	85	84	88	89	86	84	516

however, not surprising. The reason lies in the question itself (”In case of a problem I was instructed by an error message on how to solve the problem. (Please check I can not answer this question if there were no problems)”). Many subjects did not encounter any errors whilst they filled in the forms and therefore they chose ”I can not answer this question” which later resulted in a missing value. To counter a small sample size all missing values were replaced by using the Expectation-Maximization algorithm. This method of replacing missing values has been proven to be a valid and reliable method (Little & Rubin, 1987; Schafer & Graham, 2002). There are almost no differences between All Values and the EM values. See table 4 for a overview of the statistical values for all items.

Table 3: Missings Distribution Among All Questions

	Question									
	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
N	514	510	509	511	514	479	437	101	511	510
Missing Count	2	6	7	5	2	37	79	415	5	6
%	.4	1.2	1.4	1.0	.4	7.2	15.3	80.4	1.0	1.2

Item Analysis

The overall mean score is 4.77 with a standard deviation of 1.208 and a median of 5.00. The scores are not Gaussian distributed (Kolmogorov-Smirnov-Test; p<0.000) and clearly negativly skewed ($Skewness$= -1.254, SD= .035). An attempt to correct the distribution by

Table 4: Statistical Values for All Question with EM for Missings

Question	N	Mean	SD	Mode	S	K
q1	516	4.95	1.305	6	-1.440	1.464
q2	516	4.68	1.198	5	-0.976	0.588
q3	516	5.02	1.134	6	-1.292	1.281
q4	516	5.02	1.238	6	-1.377	1.360
q5	516	4.50	1.299	5	-0.722	-0.107
q6	516	4.87	1.201	6	-1.239	1.244
q7	516	3.89	1.548	5	-0.489	-0.763
q8	516	4.82	0.863	6	-1.386	2.886
q9	516	5.12	1.177	6	-1.527	1.921
q10	516	4.84	1.116	5	-1.110	1.076

using a reversed log function transformation was unsuccessful, as the transformation did not lead to a satisfying Gaussian distribution. The mean score for the questions varies between 3.89 in question 7 to 5.12 in question 9, resulting in a range of 1.23.

Discriminatory Power. Table 5 shows the discriminatory power of all questions.

Table 5: Overall Discriminatory Power of the Question

Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
q1	42.76	51.065	.581	.835
q2	43.04	50.683	.674	.827
q3	42.70	51.661	.654	.829
q4	42.70	52.957	.506	.842
q5	43.22	50.487	.619	.831
q6	42.85	53.819	.473	.845
q7	43.83	56.406	.206	.876
q8	42.90	55.988	.531	.841
q9	42.60	50.851	.677	.827
q10	42.88	49.763	.800	.817

The lowest discriminatory power lies in question 7 with .206 whereas the highest lies in question 10 with .800. Cronbach α (= .852) increases if question 7 would be deleted. On the other hand, Cronbach α would decrease if another question were to be deleted. This

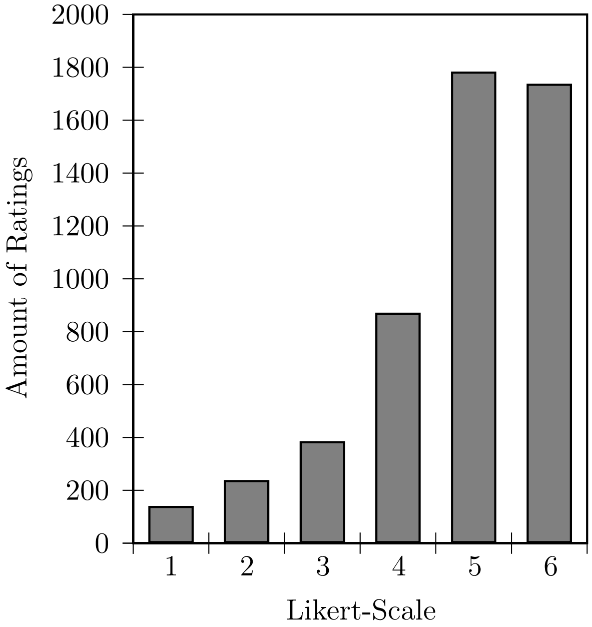


Figure 4. The Overall Distribution of the Ratings

circumstance can be viewed as evidence for good items. Table 6 shows all of the questions with their discriminatory power per form. The questions do not differ if compared to the online forms.

Table 6: Cronbach α of the Forms and Discriminatory Power of the Question per Form

		Questions, Discriminatory Power										
Form	α	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	Range
1	.804	.437	.64	.571	.593	.552	.506	.221	.531	.462	.705	.221–.705
2	.833	.363	.555	.594	.437	.570	.542	.363	.581	.655	.791	.363–.791
3	.738	.542	.454	.563	.163	.489	.505	.145	.469	.547	.604	.145–.604
4	.878	.644	.677	.599	.557	.600	.733	.432	.439	.629	.774	.432–.774
5	.871	.621	.762	.522	.615	.587	.693	.288	.573	.652	.731	.288–.762
6	.842	.568	.638	.648	.359	.585	.313	.310	.436	.706	.861	.310–.861

Correlation. Table 7 shows that all of questions indicate a medium or high correlation except for question 7 which has a poor correlation with question 1 (r= .014), 2 (r= .107), 3 (r= .034), 4 (r= .184) and 5 (r= .091). Further, question 9 and 10 have low correlations with question 7, .096 and .156 respectively. Objective data (Time and Mouse Clicks) correlate highest with question 2 (both r= -.301), question 5 (r= -.262) and question 1 (r= -.235) and amongst themselves (r= .646).

Homogeneity. All questions in the FUS should measure the construct of ”Usability” and therefore depict a positive homogeneity all throughout. The highest homogeneity is shown by questions 9 (H= .489) and 10 (H= .475) whereas the lowest is shown by question 7 (H= .228). Additionally, there are no negative correlations between the questions, indicating that they measure a similar aspect of the same construct. Homogeneity was calculated without the objective data. See table 7 for the intercorrelation matrix.

Cronbach α is high (α = .852) throughout questions, therefore it is justifiable to assume that the construct is valid.

Item Difficulty. All of the questions show a moderate range in item difficulty, ranging from .51 for question 7 with form 2 to .93 for question 9 with form 3. The mean item difficulty of all of the questions is relatively high with .795 (SD= .023) whereas the lowest mean item difficulty per question is .648 for question 7 and highest is .853 for question 9. See table 8.

Factor Analysis. A factor analysis was conducted per form and per question using Promax rotation ($Kappa$= 4) with a principal component analysis.There was no consistent evidence for factors amongst the questions on separate forms.

Table 7: Intercorrelation Matrix (Spearman, one-way) of All Question and Objective Data. Homogeneity for All Questions

Item	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	Time	Mouse Clicks
q1	1											
q2	.580	1										
q3	.568	.551	1									
q4	.300	.381	.437	1								
q5	.424	.582	.454	.341	1							
q6	.286	.294	.399	.371	.297	1						
q7	.056	.115	.080	.195	.102	.388	1					
q8	.394	.361	.222	.256	.489	.280	.353	1				
q9	.551	.517	.667	.390	.460	.376	.83	.421	1			
q10	.611	.625	.649	.365	.547	.427	.162	.688	.679	1		
Time	-.235	-.301	-.174	-.089	-.262	-.087	.079	-.072	-.118	-.202	1	
Mouse Clicks	-.170	-.301	-.167	-.108	-.186	-.170	-.045	-.055	-.113	-.147	.646	1
H	.377	.400	.402	.303	.370	.312	.228	.346	.489	.475	–	–

Table 8: Item Difficulty for all Questions and Forms.

Question	Form 1	Form 2	Form 3	Form 4	Form5	Form 6	Mean	Range	SD
q1	.894	.871	.924	.786	.831	.644	.825	.644–.924	.101
q2	.882	.784	.847	.72	.808	.636	.780	.636–.882	.089
q3	.91	.829	.916	.834	.845	.679	.836	.679–.916	.086
q4	.888	.816	.871	.863	.865	.716	.837	.716–.888	.064
q5	.835	.734	.813	.718	.775	.617	.749	.617–.835	.078
q6	.826	.737	.827	.816	.846	.819	.812	.737–.846	.038
q7	.619	.51	.572	.743	.759	.682	.648	.51–.759	.098
q8	.849	.801	.809	.823	.819	.72	.804	.72–.849	.044
q9	.905	.858	.932	.878	.859	.686	.853	.686–.932	.087
q10	.908	.82	.871	.806	.806	.623	.806	.623–.908	.098
Mean	.852	.776	.838	.799	.821	.682			
Range	.619–.91	.51–.871	.572–.932	.718–.878	.759–.865	.623–.819			
SD	.087	.104	.104	.056	.035	.060			

Discussion

The statistical values vary widely amongst the questions; due to the variation of the statistical values it is necessary to modify or delete some of the questions.

Scale

The ratings of the forms on the six-point Likert-Scale turned out to be negativly skewed. Given the fact that all of the forms created by the government, therefore having to fulfill certain requirements, it is not surprising that the scoring was above average. Furthermore, Cortesi (2008) showed that the internet presence of the Canton Basel-Stadt is above average if compared to the other Cantons in Switzerland (Basel-Stadt placed 6th out of 26). Thus, the negativly skewed distribution is explainable.

Questions

In the following part all questions will be discussed.

Question 1. The missing count for this question is one of the lowest, whilst the discriminatory power is high. The Cronbach α would decrease if this question were to be deleted. The question's homogeneity is good. Additionally, intercorrelations with the other questions and objective data are overall good with some minor discrepancies. Item difficulty for all forms is in an acceptable range. Therefore this question will not be deleted from the questionnaire.

Question 2. This question also shows a relatively low missing count. Discriminatory power is high and Cronbach α would, again, decrease if this question were to be deleted. The question's homogeneity is good. Intercorrelations are good, especially the ones with the objective data that show a medium correlation overall. Item difficulty is in an acceptable range. On this basis, this question will not be deleted from the questionnaire.

Question 3. The missing count for question 3 is low and acceptable. Discriminatory power is high. Cronbach α would decrease if this question were to be deleted. The question's homogeneity is good. The intercorrelation matrix shows some low correlations with other questions and objective data. The range of the item difficulty is acceptable. Despite the low correlations, this question will not be deleted from the questionnaire.

Question 4. A low missing count was revealed in this question. Discriminatory power was high overall, except for form number 3 (= .163). Form number 3 shows the lowest discriminatory powers with question 4 (= .163) and 7 (= .145). Cronbach α would decrease if this question were to be deleted. The question's homogeneity is good. Correlations are in mid-range whereas item difficulty is the highest amongst all of the questions. Although some statistical values could be improved, question 4 will not be excluded from the questionnaire

Question 5. Like question 1, question 5 shows the lowest missing count. Discriminatory power is high. Cronbach α would decrease if this question were to be deleted. The question's homogeneity is good. The intercorrelation matrix shows mid-range correlation with some outliers. Item difficulty is a little below average but still sufficient, however, this question will not be deleted.

Question 6. The missing value of this question is above average but still acceptable. Discriminatory power is high. Cronbach α would decrease if this question were to be deleted. The problem with question 6 is its low homogeneity (= .312) compared to the other questions. Correlations are nearly all in mid-range where objective data are the main outliers. Item difficulty shows a small range (= . 11) compared to the ranges of the other questions. However, this question will not be deleted.

Question 7. The amount of missing values is above average. Discriminatory power is low and below average across all of the forms. The item difficulty of this question differs from the others, indicating that people might not know how to answer this question. Cronbach α would increase if question 7 were to be deleted. The homogeneity of question 7 is very low (= .228). Additionally, 9 out of the 11 correlations are low. Based on these weak statistical values, question 7 will be deleted from the questionnaire.

Question 8. This question draws attention due to its high ”I can not answer this question” quote. In principle, it may seem advisable to delete question 8, however this high missing quota can be explained by the nature of the question itself. Question 8 states ”In case of a problem I was instructed by an error message how to solve the problem. (Please check I can not answer this question if there were no problems)". Many subjects did not encounter any errors whilst filling the forms in and they therefore chose "I can not answer this question" which resulted in a missing value. Unsurprisingly, there is a difference in the mean of All Values (= 4.65) and the mean with EM (= 4.82).

The discriminatory power for this question is high. Cronbach α would decrease if this question were to be deleted. Homogeneity is acceptable and correlations among the other questions are good except for the objective data. Item difficulty is in mid-range and shows a low range among the forms. Almost all of the statistical values are good and the question covers an essential part of a form, therefore this question will not be deleted.

Question 9. This questions shows a low missing count along with a high discriminatory power. Cronbach α would decrease if this question were to be deleted. The question's homogeneity is good and above average. Medium correlations can be found in the intercorrelation matrix. A broad range can be found in the item difficulty with the overall highest difficulty with form 3 (= .93). Most of the statistical values are good and so this question will not be deleted.

Question 10. A low missing count was found in this question. The discriminatory power is very high and the Cronbach α would decrease if this question were to be deleted. The question has the highest homogeneity (= .475) overall. The intercorrelation matrix shows mid-range correlations. The item difficulty shows the broadest range with this question. All of the statistical values are acceptable and thus, this question will not be deleted.

With the exception of question 7, all of the questions showed sufficient statistical values to be reused. This implies the final version of the FUS will contain only 9 questions. Results showed that the FUS is a valid and reliable questionnaire to measure usability of an online form.

To improve data, the FUS should be used and tested in a non-laboratory setting and in a real environment to figure out whether it can measure usability in a real environment too. Furthermore, the FUS should be used and tested on different kind of forms, differing in content and the frequency of which they are filled in. The FUS was, however, tested on non-interactive forms and should be further tested on interactive forms to ensure the quality of the questionnaire.

This work highlighted a necessity for a questionnaire that efficiently measures usability in online forms. With the FUS, this need is satisfied by a questionnaire that fulfills all of the requirements. One must remain aware of the fact that both forms and the world wide web are in a constant state of change and development. By this reasoning, it can not be ensured that the FUS can be used without adapting it's questions and further testing it's validity and reliability in the far future.

References

Al-Khaldi, M., & Wallace, R. (1999). The influence of attitudes on personal computer utilization among knowledge workers: the case of Saudi Arabia. *Information & management*, *36*(4), 185–204.

Bangor, A., Kortum, P., & Miller, J. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, *24*(6), 574–594.

Bargas-Avila, J., Lötscher, J., Orsini, S., & Opwis, K. (2009). Intranet satisfaction questionnaire: Development and validationof a questionnaire to measure user satisfaction with the Intranet. *Computers in Human Behavior*, *25*(6), 1241–1250.

Bargas-Avila, J., O., B., S.P., R., A.N., T., S., O., & K., O. (2010). Simple but Crucial User Interfaces in the World Wide Web: Introducing 20 Guidelines for Usable Web Form Design. *In User Interfaces, Rita Matrai (Ed.), ISBN: 978-953-307-084-1, INTECH.*

Batinic, B. (2000). *Internet für Psychologen.* Göttingen: Hogrefe.

Bevan, N. (1995). Measuring usability as quality of use. *Software Quality Journal*, *4*(2), 115–130.

Borg, I. (2002). *Mitarbeiterbefragungen-kompakt.* Göttingen: Hogrefe.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189–194.

Brynjolfsson, E., & Smith, M. (2003). Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, *49*(11), 1580–1596.

Chin, J., Diehl, V., & Norman, K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 213–218).

Cortesi, S. (2008). *ZeGo 2008-Zufriedenheit im eGovernment gemessen an den 26 Kantonsportalen der Schweiz (unpublished).*

DeLone, W., & McLean, E. (1992). Information systems success: the quest for the dependent variable. *Information systems research*, *3*(1), 60–95.

DIN, E. (1998). 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs)–Part 11: Guidance on usability. *International Organization for Standardization.*

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 345–352).

Ganassali, S. (2008). The influence of the design of web survey questionnaires on the quality of responses. In *Survey research methods* (Vol. 2, pp. 21–32).

Goodhue, D. (1993). User evaluations of MIS success: what are we really measuring? In *System* *sciences, 1992. proceedings of the twenty-fifth hawaii international conference on* (Vol. 4, pp. 303–314).

Hamilton, S., & Chervany, N. (1981). Evaluating information system effectiveness-Part I: Comparing evaluation approaches. *MIS quarterly*, 55–69.

Hassenzahl, M., Beu, A., & Burmester, M. (2001). Engineering joy. *Software, IEEE*, *18*(1), 70–76.

Herman, J. (2004). A process for creating the business case for user experience projects. In *Chi 04 extended abstracts on human factors in computing systems* (pp. 1413–1416).

Hoffmann, D., Zimmerman, S., & Tompkins, C. (1996). The dangers of directives or the false security of forms. *Journal of Law Medicine and Ethics*, *24*, 5–17.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies*, *64*(2), 79–102.

Huang, J., Yang, C., Jin, B., & Chiu, H. (2004). Measuring satisfaction with business-to-employee systems. *Computers in human behavior*, *20*(1), 17–35.

Ives, B., & Olson, M. (1984). User involvement and MIS success: a review of research. *Management Science*, *30*(5), 586–603.

James, J., Beaumont, A., Stephens, J., & Ullman, C. (2002). *Usable Forms for the Web.* Krefeld: Glasshaus.

Jarrett, C., & Gaffney, G. (2008). *Forms that work: designing web forms for usability.* Burlington: Morgan Kaufmann Pub.

Koyani, S. (2006). *Research-based web design & usability guidelines.* Washington: US General Services Administration.

Lewis, J. (1991). *User satisfaction questionnaires for usability studies: 1991 manual of directions for the ASQ and PSSUQ* (Tech. Rep.). Tech. Rep.

Lewis, J. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78.

Lindgaard, G., & Dudek, C. (2003). What is this evasive beast we call user satisfaction? *Interacting with computers*, *15*(3), 429–452.

Little, R., & Rubin, D. (1987). Statistical analysis with missing data.

Melone, N. (1990). A theoretical assessment of the user-satisfaction construct in information systems research. *Management Science*, *36*(1), 76–91.

Miller, J., & Doyle, B. (1987). Measuring the effectiveness of computer-based information systems in the financial services sector. *Mis Quarterly*, 107–124.

Mummendey, H. (1995). Die Fragebogen-Methode: Grundlagen und Anwendung in Persönlichkeits-, Einstellungs-und Selbstkonzeptforschung. 2., korr. *Aufl. Göttingen: Hogrefe, Verlag für*

Psychologie.

Newman, W., & Taylor, A. (1999). Towards a methodology employing critical parameters to deliver performance improvements in interactive systems. In *Proceedings of interact* (Vol. 99, pp. 605–612).

Niedermann, I., & Uhr, M. (2008). E-Commerce: Mehr Kunden dank Formular-Usability. *i-com*, *6*(3/2007), 55–56.

Nielsen, J. (2002). The usability engineering life cycle. *Computer*, *25*(3), 12–22.

Park, K., & Hwan Lim, C. (1999). A structured methodology for comparative evaluation of user interface designs using usability criteria and measures. *International Journal of Industrial Ergonomics*, *23*(5-6), 379–389.

Quesenbery, W. (2001). What Does Usability Mean: Looking BeyondEase of Use'. In *Annual conference-society for technical communication* (Vol. 48, pp. 432–436).

Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion.* Bern: Hogrefe & Huber.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological methods*, *7*(2), 147.

Shackel, B. (1991). Usability-context, framework, definition, design and evaluation. *Human factors for informatics usability*, 21–37.

Shirani, A., Aiken, M., & Reithel, B. (1994). A model of user information satisfaction. *ACM SIGMIS Database*, *25*(4), 17–23.

Spool, J. (1999). *Web site usability: a designer's guide.* Burlington: Morgan Kaufmann Pub.

Symons, V. (1991). A review of information systems evaluation: content, context and process. *European Journal of Information Systems*, *1*(3), 205–212.

Szajna, B., & Scamell, R. (1993). The effects of information system user expectations on their performance and perceptions. *Mis Quarterly*, 493–516.

Tullis, T., & Stetson, J. (2004). A comparison of questionnaires for assessing website usability. In *Usability professional association conference.*