Vol. 10, Issue 2, February 2015, pp. 68-86

SUPR-Q: A Comprehensive Measure of the Quality of the Website User Experience

Jeff Sauro
Principal
MeasuringU
201 Steele Street
Suite 200
Denver, Colorado
United States
jeff@measuringu.com

Abstract

A three part study conducted over five years involving 4,000 user responses to experiences with over 100 websites was analyzed to generate an eight-item questionnaire of website quality—the Standardized User Experience Percentile Rank Questionnaire (SUPR-Q). The SUPR-Q contains four factors: usability, trust, appearance, and loyalty. The factor structure was replicated across three studies with data collected both during usability tests and retrospectively in surveys. There was evidence of convergent validity with existing questionnaires, including the System Usability Scale (SUS). The overall average score was shown to have high internal consistency reliability (α = .86). An initial distribution of scores across the websites generated a database used to produce percentile ranks and make scores more meaningful to researchers and practitioners. The questionnaire can be used to generate reliable scores in benchmarking websites, and the normed scores can be used to understand how well a website scores relative to others in the database.

Keywords

usability, website quality, questionnaires, user experience

Introduction

Online consumers have many choices when making purchases or finding information on websites. If users cannot find information or purchase a product easily, they go elsewhere and may tell their friends and colleagues about the poor experience. Usability has therefore become a key differentiator for websites. Usability is a combination of effectiveness, efficiency, and satisfaction as outlined in the ISO 9241 pt. 11 definition (ISO, 1998). In practice, usability is operationalized as the combination of users' actions and attitudes. Websites are evaluated typically by observing a representative set of users attempting a set of realistic task scenarios. Users' attitudes are typically measured using a post-study and/or post-task questionnaire (Sauro & Lewis, 2009). Standardized usability questionnaires, as opposed to homegrown questionnaires, have been shown to provide a more reliable measure of usability (Hornbæk, 2006).

Standardized questionnaires alone are not particularly effective at diagnosing problems because they do not provide behavioral data. The types of questions asked are usually at too high of a level to isolate particular issues (e.g., "The website is easy to use"). However, they are one of the most efficient ways of gauging the perceived usability of an experience using measures that can most easily be compared across disparate products and domains.

Standardized usability questionnaires first appeared in the late 1980s and are widely used today (see Sauro & Lewis, 2012, Chapter 8.). Those first questionnaires were technology agnostic, meaning the items were appropriate for software, hardware, mobile devices, and websites. The advantage of a technology agnostic instrument is that the scores can be compared regardless of the technology. A company can use the same set of scores to benchmark mobile applications as well as desktop interfaces. The disadvantage of a technology agnostic instrument is that it can omit important information that is specific to an interface type.

For websites, usability is one aspect of the user experience but unlike products that are purchased and used repeatedly, the typical website experience involves other factors such as trust. The purpose of this paper is to report results of the development of a standardized questionnaire that measures several critical aspects of the website user experience. In addition, for the questionnaire to be useful, it needed to be short enough not to be a burden on participants and researchers, contain a reference database to bring more meaning to the scores, and include questions specific to the website user experience but not so specific that they are irrelevant on the disparate types of websites (e.g., non-profit versus e-commerce websites).

There are a number of published instruments that measure various aspects of website quality. Details about them (including subscales, number of items, and reliabilities) are listed in Table 1. The most commonly used instruments are technology agnostic and were developed before the web as we know it existed.

Table 1. Questionnaires That Measure Aspects of Software and Website Quality, Especially Usability With Total Number of Items and Reported Reliabilities by Overall and Subscale Constructs

Questionnaire	# Items	Measures	Global reliability	Sub-constructs	Construct reliability	Source
SUS	10	System usability	0.92	Usability	0.91	Brooke (1996)
				Learnability	0.71	Borsci et al. (2009); Sauro & Lewis (2009)
PSSUQ	16	Perceived satisfaction	0.94	System quality	0.9	Lewis (1992)
				Information quality	0.91	
				Interface quality	0.83	
SUMI	50	Usability	0.92	Efficiency	0.81	Kirakowski (1996)
				Affect	0.85	
				Helpfulness	0.83	
				Control	0.71	
				Learnability	0.82	
QUIS	27	Interaction satisfaction	0.94	Overall reaction	n/r	Chin et al. (1988)
				Screen factors	n/r	
				Terminology and system feedback	n/r	
				Learning factors	n/r	
				System capabilities	n/r	
WAMMI	20	Website usability	0.90	Attractiveness	0.64	Kirakowski & Cierlik (1998)
				Controllability	0.69	
				Efficiency	0.63	
				Helpfulness	0.70	
				Learnability	0.74	
WQ	25	Website quality	0.92	Specific content	0.94	Aladwani & Palvia (2002)
				Content quality	0.88	
				Appearance	0.88	
				Technical adequacy	0.92	
WU	8	Website usability	n/r	Ease of navigation	0.85	Wang & Senecal (2007)

Questionnaire	# Items	Measures	Global reliability	Sub-constructs	Construct reliability	Source
				Speed	0.91	
				Interactivity	0.77	
IS	15	Information satisfaction	n/r	Customer centeredness	0.92	Lascu & Clow (2008)
				Transaction reliability	0.80	
				Problem-solving ability	0.77	
				Ease of navigation	0.61	
ISQ	13	Intranet satisfaction	0.89	Content quality	0.89	Bargas-Avila et al. (2009)
				Intranet usability	0.90	
UMUX	4	Perceived usability	0.94	Perceived usability	0.94	Finstad (2010)
UMUX-LITE	2	Perceived usability	0.82	Perceived usability	0.82	Lewis et al. (2013)
HQ	7	Hedonic quality	n/r	Ergonomic quality	n/r	Hassenzahl (2001)
				Appeal	n/r	
ACSI	14-20	Customer satisfaction	n/r	Quality	n/r	theacsi.org
				Freshness of information	n/r	
				Clarity of site organization	n/r	
				Overall satisfaction	n/r	
				Loyalty	n/r	
CXi	3	Customer experience	n/r	Usefulness	n/a	forrester.com
				Usability	n/a	
				Enjoyability	n/a	
NPS	1	Customer loyalty	n/a	Customer loyalty	n/a	Reichheld (2003)
TAM	12	Technology acceptance	n/r	Usefulness	0.98	Davis (1989)
				Ease of use	0.94	
WEBQUAL	36	Website quality	n/r	Informational fit to task	0.86	Loiacono et al. (2002)
				Tailored communication	0.80	
				Trust	0.90	
				Response time	0.88	
				Ease of understanding	0.83	

Questionnaire	# Items	Measures	Global reliability	Sub-constructs	Construct reliability	Source
				Intuitive operations	0.79	
				Visual appeal	0.93	
				Innovativeness	0.87	
				Emotional appeal	0.81	
				Consistent image	0.87	
				Online completeness	0.72	
				Relative advantage	0.81	

Note. Reliability values are Cronbach alpha. n/r = not reported, n/a = not applicable.

The 10-item System Usability Scale (SUS), developed by Brooke (1996), is, perhaps, the most used questionnaire to measure perceived usability across products and websites (Sauro & Lewis, 2009). While the SUS was not published with a normative database, enough data have been collected and much of it published, that it is possible to create a set of normed scores (Sauro, 2011). A more recent scale for measuring usability is the Usability Metric for User Experience (UMUX) developed by Finstad (2010). At just four items, it is a reliable and short questionnaire. A finding also replicated in Lewis, Utesch, and Maher (2013) in which a two-item variation, called the UMUX-LITE, was also found to be reliable and correlated highly with the SUS.

Other frequently used technology agnostic instruments for measuring perceived usability include the Post Study System Usability Questionnaires (PSSUQ; Lewis, 1992), the Software Usability Measurement Inventory (SUMI; Kirakowski, 1996), and the Questionnaire for User Interaction Satisfaction (QUIS; Chin, Diehl, & Norman, 1988). The SUMI contains a reference database maintained by its authors, but at 50 items the instrument is the longest among those researched.

There are other instruments that measure factors other than usability. A standardized questionnaire to measure website quality and related constructs is the Website Analysis and Measurement Inventory (WAMMI; Kirakowski & Cierlik, 1998). The current version of the WAMMI has a set of 20 items covering the five subscales of Attractiveness, Controllability, Efficiency, Helpfulness, and Learnability and the global WAMMI measure. The WAMMI, like the SUMI, contains a reference database based on data collected from users of the questionnaire and maintained by its authors. Users of the WAMMI can convert their raw score into a percentile rank based on the scores from the other websites in the database. The internal consistency reliability of the WAMMI global score is high (ɑ =.90), whereas the subscale reliability estimates are generally lower (ɑ =.63 to ɑ =.74). The lower reliability is a tradeoff for using fewer items to measure a construct (Bobko, 2001). The WAMMI uses four items to measure each of the five constructs. Brevity is often critical when participants' time is already limited, so the loss in reliability can be justified by the higher response rates and adoption. Information about the number and type of websites in the database is not provided in the reports, but this slightly shorter multifactor instrument with a reference database was a model for the current research. The database behind WAMMI makes it appealing to generate comparison scores. The Customer Experience Index (CXi), developed by the consulting firm Forrester (www.forrester.com), is another instrument that generates comparison scores using just a few items. The CXi consists of only three items measuring usefulness, usability, and enjoyability. There is, however, no published information on the psychometric properties of the CXi.

While websites may be treated under the broader category of software, they bring the very salient elements of trust and visual appeal into consideration. Bevan (2009) argued that to encompass the overall user experience, measures of website satisfaction need to account for likability and trust. Other researchers have found that online trust is a major determinant of e-commerce success (Keeney, 1999; Pavlou & Fygensen, 2006; Suh & Han, 2003). None of the standardized usability questionnaires included a component of trust or credibility. Safar and Turner (2005) developed a psychometrically validated trust scale consisting of two factors based on an online insurance quote system. A broader examination of website trust was also conducted by Angriawan and Thakur (2008). They found that website usability, expected product performance, security, and privacy collectively explained 70% of the variance in online trust. They also found that online trust and privacy were strong predictors of consumer loyalty, which was similar to findings by Sauro (2010) and Lewis (2012).

Table 1 also lists questionnaires that focus on aspects of quality, including extensions of the Technology Acceptance Model (TAM; Davis, 1989) for the web. The WebQual questionnaire by Loiacono, Watson, and Goodhue (2002) is a more comprehensive (but longer) 36-item measure that contains subscales including trust, usability, and visual appeal. The construct of visual appeal appears in multiple questionnaires, including the WAMMI. The Web Quality (WQ) instrument by Aladwani and Palvia (2002) contains an appearance subscale, and the influential Hedonic Quality (HQ) questionnaire developed by Hassenzahl (2001) has an appeal subscale. Additional instruments focus on narrower aspects of website quality, specifically satisfaction, including questionnaires by Wang and Senecal (2007) and Lascu and Clow (2008), or with a company intranet Bargas-Avila, Lötscher, Orsini, and Opwis (2009).

Customer loyalty plays an important role in business decisions and appears as a construct in multiple questionnaires. The most popular loyalty questionnaire is the Net Promoter Score (NPS). The NPS consists of one item with an 11-point scale (0 to 10) intended to measure customer loyalty (Reichheld, 2003). Respondents are asked to rate how likely they are to recommend a friend or colleague to a product or service. Responses of 0 to 6 are considered "detractors,” 7 to 8 "passives," and 9 to 10 are "promoters." The proportion of detractors is subtracted from the proportion of promoters to create the "net" promoter score. Research conducted by Reichheld (2006) showed that the NPS was the best or second best predictor of company growth in 11 out of 14 industries. The NPS questionnaire is used widely across many industries, and benchmark data are available by third party providers. Its high adoption rate makes it a good candidate for inclusion for this current research for developing the SUPR-Q. Similar loyalty measures appear in website questionnaires from the American Customer Satisfaction Index (ACSI), maintained by the University of Michigan (www.theacsi.org) and by the company ForeSee (a proprietary instrument with no published reliabilities or details), which is used by many websites.

Based on this review, the most common constructs are measures of usability, trust, appearance, and loyalty. Some research suggests (e.g., Sauro, 2010 and Lewis, 2012) that these are overlapping constructs, as many were found to be correlated (e.g., trust and usability and usability and loyalty). These constructs formed the basis of the items used to create a new website questionnaire—the SUPR-Q.

In summary, a new instrument should be

- Generalizable: It needs to provide enough dimensions to sufficiently describe the quality of a website but not be so specific that it cannot be used on many different types of websites. For example, information websites differ from e-commerce websites which in turn differ from non-profit websites. Item phrasing needs to be generic enough so the same items can be used.
- Multidimensional: It needs to encompass the most well-defined factors for measuring website quality as uncovered in the review of existing instruments.
- Brief: It needs to be brief as time with participants is precious, and with the increase in mobile usage, makes answering lengthy questionnaires on small screens prohibitive.
- Normed: It needs to contain a normative database because knowing where a website scores relative to its peers in a normative database will provide additional information to researchers who administer the instrument in isolation.

Although some of the existing instruments share some of these aspects, most notably the generalizable and multidimensional aspects, none contain all four (i.e., generalizable, brief, covering multiple constructs including trust, and with a normative database). The purpose of this study is to develop an instrument that measures the quality of a website and is generalizable, multidimensional, brief, and backed by a normative database.

Methods

The following sections describe three studies that detail the construction of the new instrument, starting from a general approach at capturing the constructs of interest to refining the items.

Study 1

An initial set of 33 items were selected from the literature corresponding to the four constructs of usability, loyalty, trust, and appearance (based on their ability to describe website quality). The items used 5-point response options (strongly disagree = 1 to strongly agree = 5), except for the item "How likely are you to recommend the website to a friend" which used a 0 to 10-point scale. By keeping this scale format, this item can also be used to compute the Net Promoter Score (Reichheld, 2003). To assess convergent validity of the usability sub-factor, the 10 items from the System Usability Scale (SUS) were added to the survey. Tullis and Stetson (2004) found the SUS to be the best discriminating questionnaire of websites' usability.

Initial data were collected via a convenience sample in 2009. An email was sent to friends and colleagues of the author. Participants were asked to reflect on their most recent online purchasing experience and answer the 33 items plus the 10 SUS items in an online survey. A total of 100 surveys were completed that contained responses from around the US with a mix of gender (60% female, 40% male) and average age of 34 (27 to 63). Respondents were asked from which website they completed a purchase and what they purchased. Nine surveys contained at least one missing value leaving 91 fully completed surveys. In total, 51 unique websites were listed, with the most responses coming from Amazon (33), eBay (5), and Barnes & Noble (4).

An exploratory factor analysis using principal factor analysis without rotation was conducted to determine if the data was factorable and if so, how many factors to retain. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was .86 and Bartlett's Test of Sphericity was statistically significant, $\chi2$ (528) = 2191.37, $p < .001$, supporting factorability. A scree plot of the eigenvalues suggested between a three and five factor solution.

A parallel analysis was also conducted, with data showing three factors with eigenvalues greater than those from randomly simulated matrices. While the parallel analysis suggested retaining only three factors, this initial sample size was small, relative to the items being considered, and there is a theoretical rationale to look at four correlated constructs of website quality (usability, trust, appearance, and loyalty).

Given that factors were likely to be correlated, an exploratory factor analysis using principal axis factoring with oblique rotation (direct oblimin) was then conducted and four factors were retained. Items with factor loadings less than .5 were removed (this removed seven items). While the cutoff for retaining items is based somewhat on the preference of the researcher, Tabachnick and Fidell (2012) recommended a minimum .32 loading. The remaining four factors were named loyalty, trust and credibility, usability, and appearance based on the item content for items that loaded on each factor.

The remaining 26 items were broken out into their corresponding factors, and a reliability analysis was conducted for each factor. In keeping with the goal of a parsimonious instrument, items were winnowed down to as few as possible per factor. For each factor, items with item-total correlations less than .5 and with cross-loadings on multiple factors within .2 were deleted. Of the remaining items, those with the highest factor loadings and highest item-total correlation were retained, leaving 3 to 4 items per factor. A few items had negatively worded tones and those were dropped to keep an all positive instrument to avoid coding and interpretation problems (Sauro & Lewis, 2011).

The exploratory factor analysis was rerun using principal axis factoring with oblimin rotation to extract four factors. The factors, items, and communality are shown in Table 2. A total of 13 items remained.

Table 2. Item Loadings and Communalities for the 13 Remaining Items

	Usability	Trust	Loyalty	Appearance	Communality
I am able to find what I need quickly on this website.	**0.88**	0.10	-0.02	0.15	0.81
It is easy to navigate within the website.	**0.87**	-0.09	0.02	-0.05	0.78
This website is easy to use.	**0.80**	-0.13	0.09	-0.01	0.67
I feel comfortable purchasing from this website.	0.02	**-0.92**	-0.06	0.00	0.84
This website keeps the promises it makes to me.	-0.05	**-0.89**	0.05	0.04	0.80
I feel confident conducting business with this website.	0.10	**-0.89**	0.06	-0.11	0.82
I can count on the information I get on this website.	0.02	**-0.88**	-0.07	0.07	0.78
I consider myself a loyal customer of this website.	0.01	-0.03	**0.94**	-0.05	0.89
How likely are you to recommend this website to a colleague or friend?	-0.09	-0.02	**0.87**	0.12	0.78
I plan on continuing to purchase from this website in the future.	0.16	0.07	**0.81**	-0.02	0.69
I found the website to be attractive.	0.03	0.04	-0.09	**0.93**	0.87
The website has a clean and simple presentation.	0.15	-0.04	0.04	**0.78**	0.63
I enjoy using the website.	-0.06	-0.14	0.35	**0.67**	0.59
Eigenvalue	5.92	2.35	1.33	0.95	
% of Variance	45.51	18.06	10.25	7.29	
Cumulative %	45.51	63.57	73.82	81.11	

Note that to allow for reliability analysis of the subscales, it is necessary to retain a minimum of two items per factor. Thus, the shortest possible questionnaire that assesses four factors would have eight items. The internal-consistency reliability estimates and minimum inter-item correlations are shown in Table 3. All subscales showed reliabilities above .70 (Nunnally, 1978).

Table 3. Internal-Consistency Reliability Estimates (Cronbach Alpha) and Minimum Inter-Item Correlations for the Four Factor Solution from the 13 Remaining Items

	Cronbach's alpha	Minimum inter-item correlation
Appearance	.83	.60
Loyalty	.83	.58
Usability	.87	.67
Trust	.93	.70
Overall	.87	.12

To assess the convergent validity of the candidate subscales, scores on each subscale were averaged and correlated with the 10 SUS items, along with a composite score created by averaging all 13 items. The correlations are shown in Table 4.

Table 4. Correlations Between Factors, the Overall Score, and the SUS Score

	Usability	Trust	Loyalty	Appearance	Overall
Trust	0.68				
Loyalty	0.36	0.32			
Appearance	0.54	0.38	0.50		
Overall	0.77	0.77	0.73	0.74	
SUS	0.59	0.36	0.64	0.73	0.71

All correlations were statistically significantly different than zero at the p < .01 level. The usability, loyalty, and appearance factors all correlated at between r = .59 and .73 with the SUS. The overall composite score correlated at r = .71 with SUS. These medium to high correlations suggest convergent validity with the SUS. The medium to high correlations between factor average scores also confirms the correlation between factors as suggested in the literature and supports the use of an oblique rather than an orthogonal rotation. The different correlations between the factors and SUS are expected given that SUS was meant to measure only usability. However, SUS's higher correlation with the appearance factor suggests attitudes toward website appearance and usability are comingled. For further discussion on the relationship between website usability and appearance, see Tuch, Roth, Hornbæk, Opwisa, and Bargas-Avilaa (2012).

The results of the factor analysis and corresponding factor structure aren't terribly surprising. It's often the case that you get out what you put in (items to represent four constructs in and four factors out but only if you've identified four reasonably independent constructs and have done a good job of selecting items to measure them). However, the factor analysis identifies which of the original items load highest on the retained factors, have low cross-loadings on other factors, have a strong item-total correlation, and contribute to internal-reliability consistency. So it's very common to expect a certain factor structure, but it's unclear which set of items, if any, will have the desired attributes.

Study 2

A new study was run with a larger sample size per website, plus a broader range of websites, to replicate the factor structure, to reduce the number of items, and to begin a standardization dataset. In 2010, participants were recruited using online advertisements, Amazon's Mechanical Turk, and panel agencies to participate in a short online survey. Participants were paid between $.40 and $3 to complete the survey. Participants were from the US and were asked to attempt one predefined task on one of 40 websites and answer the 13 candidate items selected from the pretest.

The websites selected represented a range of quality. Some had poor usability by their identification on the website, webpagesthatsuck.com. Other websites were some of the most visited websites in the US. They came from a range of industries, including retail, travel, IT, government, and cellular service carriers. For example, we included websites from the NY State Government, Tally-Ho Uniforms, author Julie Garwood, Crumpler (a bag and luggage company), Expedia, Sprint, Target, Walmart, and Budget (a car rental company). The tasks participants attempted were tailored for each website (e.g., finding airline ticket prices, locations of government offices, or product prices). Task completion rates varied between 20% and 100%.

To further assess the convergent validity of the instrument, the 20 items from the WAMMI questionnaire were used in eight websites' surveys and the 10 SUS items were also included along with the 13 candidate items.

Participants completed a total of 484 surveys (each participant completed one survey). Each website had between 10 and 15 participants attempting a task on the websites. The participants were a mix of gender (58% female, 42% male), had a mix of occupations (including professionals, homemakers, and students), had a mix of education levels (45% bachelor's degree, 34% high school/GED degree, and 18% advanced degree), and represented 47 states. The median age of the participants was 33 (18–68). Participants had a range of experience with each website, with the lesser known poor-quality websites having no users who had prior experience, compared with moderate exposure for some participants on the higher-traffic websites.

To assess the factor structure, principal axis factoring using oblimin rotation was conducted. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was .92 and Bartlett's Test of Sphericity was statistically significant, χ2 (78) = 4015.20; p < .001. The factor loadings and communalities are shown in Table 5.

Table 5. Factor Loadings for the Rotated Factor Solution for the 13 Candidate Items

	Trust	Usability	Appearance	Loyalty	Communality
I can count on the information I get on this website.	**0.95**	0.03	0.05	-0.12	0.91
I feel confident conducting business with this website.	**0.73**	-0.18	-0.01	0.08	0.56
This website keeps the promises it makes to me.	**0.73**	-0.04	0.04	-0.01	0.53
I feel comfortable purchasing from this website.	**0.59**	-0.23	-0.09	0.20	0.45
The information on this website is valuable.	**0.48**	0.10	0.06	0.17	0.27
I am able to find what I need quickly.	0.00	**-0.87**	0.00	0.07	0.76
This website is easy to use.	0.09	**-0.84**	0.01	0.06	0.71
It is easy to navigate within the website.	0.05	**-0.74**	0.24	-0.01	0.61
I found the website to be attractive.	0.08	0.06	**0.71**	0.14	0.54
The website has a clean and simple presentation.	0.04	-0.29	**0.67**	-0.03	0.53
I will likely purchase something from this website in the future.	0.01	-0.01	0.06	**0.69**	0.48
How likely are you to recommend the website to a friend or colleague?	0.14	-0.19	0.07	**0.57**	0.39
I enjoy using the website.	0.09	-0.23	0.25	**0.40**	0.28
Extraction sums of squared loadings	7.45	0.88	0.49	0.32	
% of Variance	57.33	6.79	3.78	2.46	
Cumulative %	57.33	64.11	67.89	70.36	
Rotation sums of squared loadings	5.92	5.52	4.90	4.90	

In examining the factor loadings in Table 5, the items still fit a four-factor structure reasonably well with most loadings above .6. To further reduce the number of items, two items "I enjoy using the website" and "The information on this website is valuable" had the lowest loading on their respective factors and were dropped.

To assess the convergent validity of the four subscales, scores were created by averaging the item scores for each subscale and correlating them with SUS (n = 441) and WAMMI (n = 106). The correlations are shown in Table 6.

Table 6. Correlations Between Subscales and the SUS and WAMMI Questionnaires

	Usability	**Trust**	**Loyalty**	**Appearance**	**Overall**	**SUS**
Trust	0.66					
Loyalty	0.64	0.67				
Appearance	0.68	0.64	0.63			
Overall	0.87	0.87	0.88	0.82		
SUS	0.88	0.71	0.69	0.73	0.87	
WAMMI	0.86	0.71	0.66	0.67	0.85	0.95

All correlations were statistically significant at p < .01. The usability score and overall score showed the highest convergent validity with strong correlations with both the SUS (r ≥ .87) and WAMMI (r ≥ .85). All subscales were, however, significantly and moderately to strongly correlated with both the SUS and WAMMI.

The internal consistency reliability estimates for each subscale and minimum inter-item correlations are shown in Table 7. All subscales and the overall scale show reliabilities to be above .70 except for the loyalty factor with coefficient alpha of .63. The measure created is called the Standardized User Experience Percentile Rank Questionnaire (SUPR-Q).

Table 7. Internal Consistency Reliability Estimates (Cronbach Alpha) and Minimum Inter-Item Correlations for the Four-Factor Solution From the 13 Remaining Items

	Cronbach alpha	**Min. inter-item correlation**
Appearance	.82	.69
Loyalty	.63	.61
Usability	.94	.85
Trust	.89	.44
Overall	.91	.39

For eight websites, the SUS, WAMMI, and SUPR-Q were collected for 108 total responses. A one-way analysis of variance was used, with the SUPR-Q total score as the dependent variable and the website as the independent variable with eight levels. A significant effect was found for the different websites, F(7, 100) = 4.43, p < .001 (Adj r-square = 18.34%). It exhibited the same or equal discriminating power as the SUS, F(7, 100) = 4.523; p < .001 (Adj r-square = 18.72%), and WAMMI, F(7, 100) = 4.22; p < .001 (Adj r-square = 17.40%). The average of the three items on the usability factor showed high discrimination, F(7, 100) 5.19; p < .001 (Adj r-square = 21.52%), as did the loyalty subscale, F(7, 100) 5.57; p < .001 (Adj r-square = 23.01%). To a lesser extent, trust discriminated, F(7, 100) 2.91; p = .008 (Adj r-square = 11.12%), and appearance did not significantly discriminate, F(7, 100) 1.39; p = .22.

The 13 items loaded as expected on a four-factor structure. Two of the items had low loadings and were dropped. The remaining 11 items showed high overall internal consistency reliability and sensitivity in discriminating between websites with poor and high quality. The reliability of the subscales was high, with the exception of the loyalty factor, which had a coefficient alpha below .70, below the generally acceptable cutoff (Nunnally, 1978). There are two possible reasons for the low internal consistency reliability. The first possible reason is the potential non-relevance for the item "I will likely purchase something from this website in the future." For some websites used in the study, participants couldn't make a purchase, rendering the item irrelevant. A possible work-around is to make the item wording more generic "I will likely visit this website again in the future." This wording will be subsequently tested in future studies for information sites that don't support purchasing. The second reason is that the loyalty factor uses an 11-point scale to keep the scoring consistent with industry practices. Correlations will be attenuated when items use a different number of response options—this in turn reduces the estimates of internal consistency reliability. The loyalty factor has only two items and the tradeoff for fewer items is lower reliability, which in this case, while low, it's similar to that found for factors on the WAMMI.

Finally, a few participants in the survey comments noted that the item "This website keeps the promises it makes to me" sounded awkward. One participant commented for example, "how can a website keep promises?" Rewording the item to "The website keeps the promises it makes" was tested in the subsequent study.

Study 3

A third study was conducted with a larger sample size to test the new wording of items and to further reduce the items on the trust and usability factors that had more than two items each. Between 2011 and 2012, participants were recruited using online advertisements, Mechanical Turk, and panel agencies to participate in a short online survey. To qualify, participants must have visited or made a purchase on one of 70 websites in the prior six months. Participants were paid between $.40 and $2.50 to complete the survey. Unlike Study 2 where participants were randomly assigned to different websites and asked to complete a task, this study asked participants to reflect on their recent website experience based on whatever interaction they had. Participants were only allowed to respond to one website, even if they qualified for multiple websites. This ensured that the responses were not relative comparisons, which can skew the ratings.

Participants completed a total of 3,891 surveys (each participant completed one survey for one website). The participants were a mix of gender (53% female, 47% male), had a mix of occupations (including professionals, homemakers, and students), and had a mix of education levels (46% bachelor’s degree, 40% high school/GED degree, and 5% advanced degree). The median age of the participants was 29 (18–73). Participants had a range of experience with each website, with each participant having had to visit the website at least once in the prior six months. Participants reflected on their experience with 51 websites across a range of industries, including companies and institutions such as Delta, Craigslist, USA.gov, eBay, New York Times, Office Depot, and Walgreens. The median number of responses per website was 44 (14–126 responses).

To assess the factor structure with this new set of data, a principal axis factoring using oblique rotation was conducted. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was .93 and the Bartlett's Test of Sphericity was statistically significant, χ2 (55) = 26897.4, p < .001. The factor matrix is shown in Table 8.

Table 8. Factor Loadings for the Rotated Factor Solution for the 11 Candidate Items

	Usability	**Trust**	**Loyalty**	**Appearance**
It is easy to navigate within the website.	**0.88**	0.01	0.00	0.00
The website is easy to use.	**0.87**	0.01	0.02	-0.01
I am able to find what I need quickly on the website.	**0.58**	0.09	0.12	0.11
I feel comfortable purchasing from the website.	0.02	**0.86**	-0.07	0.01
I feel confident conducting business on the website.	0.06	**0.84**	0.05	-0.04
I can count on the information I get on the website.	-0.01	0.35	0.31	0.20
I will likely return to the website in the future.	0.05	-0.01	**0.78**	-0.03
How likely are you to recommend the website to a friend or colleague?	0.04	-0.02	**0.77**	0.04
The website keeps the promises it makes to me.	-0.01	0.35	0.39	0.16
I find the website to be attractive.	-0.01	0.01	0.03	**0.76**
The website has a clean and simple presentation.	0.34	-0.01	-0.03	**0.56**

	Usability	Trust	Loyalty	Appearance
Extraction sums of squared loadings	5.90	0.91	0.44	0.20
% of Variance	53.67	8.25	3.97	1.80
Cumulative %	53.67	61.92	65.88	67.68
Rotation sums of squared loadings	4.82	3.89	4.54	4.62

Three items were dropped. The item "I am able to find what I need quickly on the website" had the lowest relative loading on the usability factor and was dropped. The item "The website keeps the promises it makes to me" and "I can count on the information I get on the website" both had loadings below .4 and cross-loaded on multiple factors. This reduced the total number of items to eight.

To assess the factor structure with these eight items, principal axis factoring using oblique rotation was conducted. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was .86 and the Bartlett's Test of Sphericity was significant, χ2 (28) = 17512, p < .001. The final factor matrix is shown in Table 9.

Table 9. Factor Loadings for the Rotated Factor Solution for the Eight Remaining Items

	Usability	Trust	Loyalty	Appearance
The website is easy to use.	**0.88**	0.02	0.02	-0.02
It is easy to navigate within the website.	**0.80**	0.02	0.03	0.06
I feel comfortable purchasing from the website.	-0.01	**0.87**	-0.05	0.02
I feel confident conducting business on the website.	0.03	**0.83**	0.08	-0.02
How likely are you to recommend the website to a friend or colleague?	-0.01	-0.01	**0.80**	0.05
I will likely return to the website in the future.	0.03	0.01	**0.79**	-0.03
I find the website to be attractive.	-0.05	0.03	0.05	**0.76**
The website has a clean and simple presentation.	0.25	0.00	-0.02	**0.64**
Extraction sums of squared loadings	4.26	0.80	0.42	0.18
% of Variance	53.24	10.05	5.30	2.26
Cumulative %	53.24	63.30	68.60	70.85
Rotation sums of squared loadings	3.53	2.77	3.26	3.47

The final eight-item SUPR-Q reflects the multi-factor solution for measuring the quality of the user experience of websites and having a normalized database with over 100 websites.

To assess the convergent validity of the eight-item SUPR-Q, scores were created by averaging the items for the global score and for each subscale for each participant's response (participant level scoring). For a subset of the websites, the 10-item SUS was also collected and the global and subscales were correlated at the participant level (n = 2,513). The correlations are shown in Table 10 below.

Table 10. Correlations Between Subscales, Overall Score and the SUS Done at the Individual Response Level

	SUS	SUPR-Q	Usability	Trust	Loyalty
SUPR-Q	0.75				
Usability	0.73	0.85			
Trust	0.39	0.62	0.46		
Loyalty	0.61	0.84	0.60	0.49	
Appearance	0.64	0.85	0.73	0.48	0.57

All correlations calculated at the participant level were statistically significantly different than zero (p < .001). The usability factor score and overall score showed high convergent validity with strong correlations with SUS (r ≥ .73).

Correlations were calculated again at the study level (study level coding) where the average score for each website was correlated (n = 40). It has been shown that study level metrics tend to correlate higher than individual metrics as the variance within studies is eliminated (Sauro & Lewis, 2009), and it is the study-level scores that are of interest to researchers. The correlations are shown in Table 11.

Table 11. Correlations Between Subscales, Overall Score and the SUS Done at the Study Level (Averaged Across Respondent by Website)

	SUS	SUPRQ	Usability	Trust	Loyalty
SUPRQ	0.87				
Usability	0.87	0.88			
Trust	0.47	0.57	0.40		
Loyalty	0.82	0.91	0.73	0.72	
Appearance	0.73	0.86	0.81	0.71	0.64

All correlations at the study level were statistically significantly different from zero (p < .001). The usability factor score and overall score showed high convergent validity with strong correlations with SUS (r = .87). A reliability analysis was conducted on the four factors, and the coefficient alpha and item correlations are shown in Table 12.

The overall composite score made up of eight items and the usability factors both have high reliability (coefficient alpha > .85), and the trust and appearance factors have acceptable reliability (coefficient alpha > .75), with the loyalty factor having low reliability (coefficient alpha = .64).

Table 12. Internal-Consistency Reliability Estimates (Cronbach Alpha) and Minimum Inter-Item Correlations for the Four Factor Solution From the Eight Remaining Items

	Cronbach alpha	Min. / inter-item correlation
Appearance	.78	.64
Loyalty	.64	.65
Usability	.88	.78
Trust	.85	.73
Overall	.86	.36

For 40 websites, the SUS and eight candidate items were collected for 2,513 total responses. A one-way ANOVA was used with the combined average score for the eight items and website as the independent variable with 40 levels. The combined average discriminated well between the poorest and highest quality websites, F(39, 2473) = 10.22; p < .001 (Adj r-square = 12.52%). It exhibited about equal discriminating power as the SUS, F(39, 2473) = 9.67; p < .001 (Adj r-square = 11.86%) with two fewer items. The subfactors also provided evidence for sensitivity by discriminating between the websites on

- usability, F(39, 2473) = 6.03; p < .001 (Adj r-square = 7.25%);
- trust, F(39, 2473) = 12.13; p < .001 (Adj r-square = 14.73%);
- loyalty, F(39, 2473) = 14.80; p < .001 (Adj r-square = 17.65%); and
- appearance, F(39, 2473) = 5.82; p < .001 (Adj r-square = 6.96%).

The distribution of the average scores for the overall composite and the subscales for the 70 websites is shown in the histograms in Figure 1.

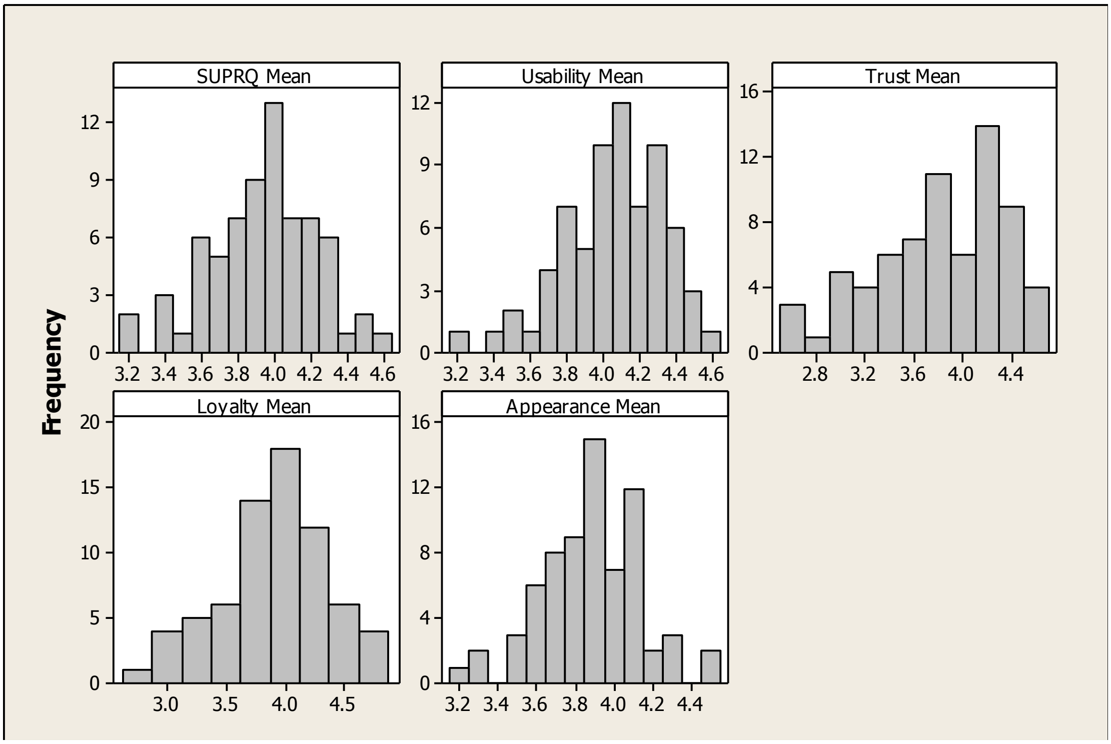

Figure 1. Distribution of subscales and overall score for 2,513 responses across 70 websites.

The distribution of scores is generally normally distributed, with a skewness for appearance of -.10 and skewness for usability of -.56; all other skewness values were intermediate to these two. The means and standard deviations for each of the subscales and overall are shown in Table 13.

Table 13. Mean and Standard Deviations for 2,513 Responses Across 70 Websites by Overall Score and Subscale Scores

	Mean	**Standard Deviation**
SUPRQ	3.93	0.29
Usability	4.06	0.29
Trust	3.80	0.52
Loyalty	3.91	0.46
Appearance	3.88	0.25

The means and standard deviations can be used as the basis for identifying percentile ranks for the overall score and subscale scoring. For example, a website that obtains an overall mean score of 4.1 would be about .6 standard deviations above the mean. This would place it higher than 70% of websites in the database. Its score can then be expressed as a 70, meaning a percentile rank of 70. While the shape of the distributions are reasonably normal, additional data may skew the values more or reduce skewness. Future analysis will need to examine the distributions to determine if a log-transformation is needed to maintain a normal distribution. This is essential for computing accurate normed scores.

Discussion

The goals of this study were to develop an instrument that measures the quality of the website user experience that is short but still reliable, and to begin to construct a normative database.

The SUPR-Q measures four aspects of the quality of the user experience: usability, trust, appearance, and loyalty. It does not have subscales to measure aspects that were included in the literature review such as assessment of the utility of features and functions or the emotional/hedonic appeal of websites. Future analysis should investigate how much these attributes correlate with the existing factors in the SUPR-Q. Future analysis should also investigate additional item phrasing to avoid using the word "business" on the credibility factor (from the item "I feel confident conducting business with this website”). There are a number of information-only websites, educational and non-profit websites, and corporate websites where conducting business is not relevant or in some cases is illegal (e.g., direct selling from drug manufacturing websites). The data suggest, however, that despite the labeling of "business" the responses to the websites included in these studies still provide a valid and reliable measure.

The number of items was reduced to two per factor—the minimum number possible to still perform a reliability analysis. The consequence of reducing the number of items overall and per factor was a reduction in reliability. Although lower in reliability, the reliability is still sufficiently high for two items to make the instrument useful. The lowest reliability is on the loyalty factor. This is partly a consequence of mixing an 11-point item with a 5-point item, which attenuates the correlation and consequently, the reliability. Future analysis should investigate whether a third item can increase the reliability of this factor. With the lower reliability, practitioners should plan on using larger sample sizes, especially if it's critically important to get an accurate measure of loyalty. Enough data were collected in these initial studies to generate a relative distribution of scores to provide percentile rankings. Future studies can be conducted to add to the database and examine wider ranges of websites.

The high correlation between usability and appearance across the studies suggests a comingled relationship. This strong correlation was seen with both the usability factor of the SUPR-Q and the SUS. It suggests that participants are rating more attractive websites more usable (or vice-versa). Future analysis should continue to examine the relationship between appearance and usability, similar to the Tuch et al. (2012) paper.

Conclusion

Over 4,000 responses to experiences with over 100 websites were analyzed to generate an eight-item measure of website quality. The questionnaire is called the Standardized User Experience Percentile Rank Questionnaire (SUPR-Q) and contains four factors: usability, trust, appearance, and loyalty.

The factor structure was replicated across three studies with data collected both during a usability test and retrospectively. There was evidence of convergent validity with existing questionnaires SUS and WAMMI. The overall average score was shown to have high internal consistency reliability (ɑ = .86), while the subscales had lower but generally acceptable levels of reliability (ɑ = .64 to ɑ= .88). The lower reliability is a consequence of using only two items per factor to keep the total length short—one of the primary goals of this research. Finally, an initial distribution of scores across the websites generated a database to generate percentile ranks and make scores more meaningful to researchers.

To administer the SUPR-Q, users responded to seven of the eight items using a 5-point scale (1 = strongly disagree and 5 = strongly agree). For one item ("How likely are you to recommend this website to a friend or colleague?") users respond to an 11-point scale (0 = not at all likely and 10 = extremely likely). The following are the eight items in the SUPR-Q and their corresponding factor:

- The website is easy to use. (usability)
- It is easy to navigate within the website. (usability)
- I feel comfortable purchasing from the website. (trust)
- I feel confident conducting business on the website. (trust)
- How likely are you to recommend this website to a friend or colleague? (loyalty)
- I will likely return to the website in the future. (loyalty)
- I find the website to be attractive. (appearance)
- The website has a clean and simple presentation. (appearance)

Tips for Usability Practitioners

When assessing the quality of the website user experience, practitioners should consider using a standardized questionnaire.

- Standardized questionnaires provide a more reliable and valid measure of the construct of interest compared to homegrown questionnaires.
- Standardized questionnaires, like the SUPR-Q, can be administered during a usability test or outside of a usability test. However, they are not a substitute for usability testing as they can't pinpoint problems in an interface.
    - The SUPR-Q can be administered before and after website changes to measure how much improvement, if any, was achieved.
    - The SUPR-Q normed scores allow website owners the ability to benchmark their website's usability, trust, appearance, and loyalty.
- Practitioners should consider using a standardized questionnaire that measures more than just a single factor such as usability. While usability is a critical component of website quality, it does not fully describe the entire website user experience.

Acknowledgements

Thanks to Kathy Green PhD and Jim Lewis PhD for providing suggestions on earlier drafts of this paper.

References

Aladwani, A. M., & Palvia, P. C. (2002). Developing and validating an instrument for measuring user perceived web quality. *Information Management, 39*, 467–476.

Angriawan, A., & Thakur, R. (2008). A parsimonious model of the antecedents and consequence of online trust: An uncertainty perspective. *Journal of Internet Commerce, 7*(1) 74–94.

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

Bevan, N. (2009). Extending quality in use to provide a framework for usability measurement. In *Human Centered Design* (pp. 13–22). Heidelberg, Germany: Springer Berlin.

Bobko, P. (2001). *Correlation and regression: Applications for industrial organizational psychology and management*. New York: Sage Publications.

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the system usability scale: A test of alternative measurement models. *Cognitive Processes, 10*, 193–197.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 189–194). London, UK: Taylor & Francis.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human–computer interface. In *Proceedings of CHI 1988* (pp. 213–218). Washington, DC: ACM.

Davis, D., (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13(3)*, 319–339.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327.

Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human Computer Interaction, 13(4)*, 481–499.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64(2)*, 79–102.

ISO, 1998 (1998). *Ergonomic requirements for office work with visual display terminals (VDTs), Part 11, Guidance on usability* (ISO 9241-11:1998E), Geneva, Switzerland: ISO.

Keeney, R. L. (1999). The value of internet commerce to the customers. *Management Science*, *45*(4), 533–542.

Kirakowski, J. (1996). The Software Usability Measurement Inventory: Background and usage. In P. Jordan, B. Thomas, B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 169–178). London, UK: Taylor & Francis.

Kirakowski, J., & Cierlik, B. (1998). Measuring the usability of websites. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 424–428). Santa Monica, CA: HFES. Also available at www.wammi.com (accessed April 15, 2014).

Lascu, D., & Clow, K. E. (2008). Web site interaction satisfaction: Scale development consideration. *Journal of Internet Commerce*. *7(3)*, 359–378.

Lewis, J. R. (1992). Psychometric evaluation of the Post-Study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1259–1263). Santa Monica, CA: Human Factors Society.

Lewis, J. R. (2012). *Predicting Net Promoter scores from System Usability Scale scores*. Available at www.measuringusability.com/blog/nps-sus.php (accessed April 4, 2014).

Lewis, J., Utesch, B., & Maher, D. (2013). UMUX-LITE: When there's no time for the SUS. In *Proceedings of the Conference in Human Factors in Computing Systems* (CHI 2013; pp. 2099–2102). New York, NY: ACM.

Loiacono, E. T., Watson, R. T., & Goodhue, D. L. (2002). WEBQUAL: A measure of website quality. *Marketing Theory and Applications*, *13(3)*, 432-438.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Pavlou, P. A., & Fygenson, M. (2006). Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior. *MIS Quarterly*, *30*(1), 115–143.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*, 46–54.

Reichheld, F. (2006). *The ultimate question: Driving good profits and true growth*. Boston, MA: Harvard Business School Press.

Safar, J. A., & Turner, C. W. (2005). Validation of a two factor structure of system trust. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 497–501). Santa Monica, CA: HFES.

Sauro, J. (2010). *Does better usability increase customer loyalty*? Available at www.measuringusability.com/usability-loyalty.php (accessed April 1, 2014).

Sauro, J. (2011). *A practical guide to the System Usability Scale (SUS): Background, benchmarks & best practices*. Denver, CO: Measuring Usability LLC.

Sauro, J., & Lewis J. R. (2009). Correlations among prototypical usability metrics: evidence for the construct of usability. In *Proceedings of the Conference in Human Factors in Computing Systems* (CHI 2009; pp. 1609–1618). Boston, MA: ACM.

Sauro, J., & Lewis J. R. (2011). When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the Conference in Human Factors in Computing Systems* (CHI 2011; pp. 2215–2223). Vancouver, BC, Canada: ACM.

Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Waltham, MA: Morgan Kaufmann.

Suh, B., & Han, I. (2003). The impact of customer trust and perception of security control on the acceptance of electronic commerce. *International Journal of Electronic Commerce*, *7*(3), 135–161.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using Multivariate Statistics* (6th Edition). Boston, MA: Allyn and Bacon.

Tuch, A., Roth, S., Hornbæk, K., Opwisa, K., & Bargas-Avilaa, J. (2012). Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior*, *28*(5), 1596–1607

Tullis, T. S., & Stetson, J. N. (2004). A comparison of questionnaires for assessing website usability. Paper presented at the *Usability Professionals Association Annual Conference* (pp. 1–12). Minneapolis, MN: UPA. Available also at home.comcast.net/~tomtullis/publications/UPA2004TullisStetson.pdf (accessed October 1, 2011).

Wang, J., & Senecal, S. (2007). Measuring perceived website usability. *Journal of Internet Commerce*, *6*(4), 97–112.

About the Author

Jeff Sauro
Mr. Sauro is the Principal and founder of Measuring U (measuringu.com). He is author of five books, including *Quantifying the User Experience* and *Customer Analytics for Dummies*. He has worked for Oracle, PeopleSoft, Intuit, and General Electric and holds a Masters from Stanford in Learning, Design & Technology and is completing his PhD in Research Methods & Statistics at the University of Denver.