Association for Information Systems
AIS Electronic Library (AISeL)

Wirtschaftsinformatik Proceedings 1999 Wirtschaftsinformatik

February 1999

Satisfaction with a Web-Site: Its Measurement, Factors and Correlates

Paul Alpar
Universität Marburg, alpar@wiwi.uni-marburg.de

Satisfaction with a Web Site: Its Measurement, Factors and Correlates

Paul Alpar
Universität Marburg (alpar@mailer.uni-marburg.de)

Contents

Electronic Business Engineering / 4. Internationale Tagung Wirtschaftsinformatik 1999.
Hrsg.: August-Wilhelm Scheer; Markus Nüttgens. – Heidelberg: Physica-Verlag, 1999

Abstract

The number of Internet users as well as the number of web sites on the Internet are rapidly rising. Users have numerous web sites to choose from while surfing the net, so the web sites are competing for the users' attention and repeated visits. To achieve their goals whatever they may be the web sites must satisfy their visitors' information and communication needs as much as possible within the given economic restrictions. We have developed an instrument to measure how satisfied visitors of a web site are with the site. The instrument has been used to evaluate sites and to explore factors and correlates of user satisfaction with a web site (USW). Since Spring 1997 it has been employed to evaluate over 20 web sites by over 1000 users.

1 Introduction

The number of companies building a presence on the Internet is continuously growing. Based on a representative survey it has been estimated that in 1997 over 100,000 companies in Germany with more than 9 employees maintained a web site (Alpar 1997). According to another though not necessarily representative survey reported in the same paper the primary purposes of German firms for the establishment of a presence on the Web were to enhance their image, achieve competitive advantage, prevent competitive disadvantage, and reach new customers. When questioned how they measure the success of their Internet activities a number of firms responded that they do not measure it. The majority of those firms that do measure something use measures based on hits, page views, or similar statistics derived from server logs. Some researchers suggest indeed that this is the way to measure the success of a web site (Jaspersen 1996). Firms also observe duration of visits, the path taken through a web site, and, if possible, specific visitors' frequency of visits.

Information derived from server logs is certainly important information that needs to be tracked. However, many problems are associated with these measures. The accuracy of some of the measures can be doubted (e.g., the measurement of page view duration). Measures are often defined in different ways (e.g., there is no agreement as to what constitutes a single visit to a site) and therefore not comparable. Further, if a web site should be enhanced these measures are of limited help in deciding which issues need to be addressed. Visitors may be communicating their (dis)satisfaction and improvement suggestions by electronic mail (e-mail). This source of data is valuable but it is not a way to collect reliable data that can be systematically tracked over time. Therefore, we suggest a more direct way: to explicitly question the visitors how they like a web site, just like customers are surveyed about their satisfaction with a product or a service.

We consider a web site to be a service that provides a user with information and communication. This is true whether the user pays for the right to visit it or not. The user occurs at least opportunity costs (in terms of time spent surfing) but possibly also telecommunication and Internet access costs. He will reject or use less of a service which value to him is below his costs. More exactly, we consider a web site to be an information system service. This has relevance for the construction of the measurement instrument.

2 Related Research

The measurement of user satisfaction with a web site is an evaluation exercise. Since evaluation of web sites can be done from many different perspectives and in many different ways a variety of efforts have been undertaken under this topic. The most simple type of evaluation of web sites is found in many popular computing journals and even in some general interest publications: it is the evaluation by individual ”experts” who are often journalists. These experts give a short description of a web site and rate it by assigning it an overall value. The value is expressed in number of stars, number of flies, coloring (with the degree of redness indicating how ”hot” the site is), with an arrow or thumb (where the direction of the pointer indicates how good the site is and ”up” stands for the best evaluation), or in a similar way. In some cases the sites are evaluated using a few criteria rather than assigning them just one value. Even if the evaluators are knowledgeable such evaluations are not reliable and they are too crude to help much in enhancing a site or benchmarking several sites.

One perspective of web site evaluation is the issue of usability (Boling 1995). The approach sketched in that paper has been used during a redesign of the home page of the Indiana University Bloomington. Although the paper suggests that usability testing should go beyond design issues most of the testing concentrated on how to get the visitors from the home page to where they want to go rather than on the content of the site. This emphasis may have been appropriate in the particular case since a university home page mainly serves as a sign post to other sites.

Trochim (1996) briefly describes various approaches to the evaluation of web sites (including standard server log analyses). The author and his students applied some of them to measure the effectiveness of a course web site. The web site was developed to support courses on research design and also for experimental purposes. The evaluation bases on the idea that a web site evolves through the phases of conceptualization of the content domain, the development of the content, implementation, and evaluation and that evaluation should take place in each of these phases. Several evaluation questions are proposed for each phase. However, the evaluation questions offered for the first two phases are some of the basic questions usually not considered to be specifically evaluation questions (e.g., What are the purposes of the site? and What are the site start-up costs?).

Following issues have been analyzed: student statements about the web site as a learning tool (Fitzelle Jr./Trochim 1996), content of e-mail messages and bulletin board entries regarding the supported course (Bonanno 1996), student perceptions about the contribution of the web site to their learning process (Trochim/Hover 1996), the impact of web site use on the actual success in exams and term papers (Bowen 1996). The prevalent analysis method used was concept mapping. All the reported evaluations require a close contact with evaluators (e.g., brainstorming sessions or examinations of evaluators) which is not given in the context of many commercial web sites. However, the evaluation of web sites used for advertising and commerce was explicitly not addressed in the work by Trochim and his students.

Ho (1997) evaluated 1,800 commercial web sites from various countries and continents. He assumes in his paper that the business *purposes* of a site can be classified into three categories (promotion, provision, and processing) and that sites create *value* for their visitors in four ways (timely, custom, logistic, and sensational). The evaluation consisted of counting which and how many of the 12 possible features (purpose-value combinations) within the given framework a site offers. Then the data were aggregated for industries, countries, and regions. The reliability of such an evaluation cannot be examined and the results are of little help to individual sites. If the evaluation process would be made reliable such an evaluation could serve as a snapshot of web use by industries and geographical units.

Another effort explicitly targets web sites for electronic commerce (Selz/Schubert 1997). The whole sales transaction process supported by a web site is under consideration in this work. The transaction process is considered to consist of four phases: information, agreement, settlement, and community. The last phase refers to communications among the customers, communications between the customers and the firm owning the web site, and the building of a community. In addition, in each of the phases a firm makes different related offerings called modules: product/service, bundling, generic services, customer profile, community. For each ”most important” module in each phase some criteria have been chosen to be rated by site evaluators. However, in a related document whole phases rather than modules are rated (no author 1998). The ratings can range from 1 to 4 while a 0 denotes ”not applicable” (in the related document the range is given with –2 to 2). The ratings of criteria are averaged across raters and these averages are averaged to give the overall rating for a module or a phase. This is a very simple information aggregation method. There is absolutely no information given about the reliability or validity of the instrument or that attempts have been made to evaluate these necessary properties of a survey instrument.

Lohse/Spiller (1998) explore the effects of user interface design on traffic and sales in electronic shopping though without directly involving users in their research. They compare 32 design and site characteristics (e.g., number of products offered) of 28 online retail stores with visits and sales in one month. This approach does not really try to explain user preferences and the process of their formation since user opinions are excluded. It is also questionable whether 28 observations are enough to determine the impact of 32 features. Findings like ”the number of products in a store explains 17% of variance in store traffic but had no effect on sales” beg further investigation. Finally, if the impact of some features on sales is examined then the cost of providing these features should be considered too to make the findings more useful. Jarvenpaa/Todd (1977) analyzed user reactions to electronic shops giving qualitative insights on the topic. Table 1 summarizes some characteristics of the reported research.

Reference	**Evaluators**	**Data gathering**	**Evaluation goal**
Boling 1995	Students, parents, faculty, and staff	Observation and usage logs	Usability
Fitzelle Jr. / Trochim 1996	Students	Questionnaire	Effectiveness
Bonanno 1996	Students	Bulletin board and e-mail messages	Effectiveness
Trochim / Hover 1996	Students	Brainstorming	Effectiveness
Bowen 1996	No explicit evaluation	Usage logs and exam results	Effectiveness
Ho 1997	“Experts”	Counting of site features	Feature richness
Selz / Schubert 1997	Not specified	Questionnaire	Sales transaction support
Lohse / Spiller (1998)	No explicit evaluation	Inspection of on-line store features and usage logs	User interface design

Table 1: A Summary of Related Research

The work reported below attempts an evaluation of commercial web sites that goes beyond design issues but does not comprise whole sales transaction processes (if such transactions are offered at all). Major efforts have been expended to develop a valid and reliable instrument.

3 Measurement Instrument

The development of a reliable and valid instrument for the measurement of an abstract concept like satisfaction requires several steps. First, the need for such an instrument needs to be made plausible. This has been done in the introduction. Second, a constitutive and an operational definition need to be given.

A constitutive definition explains a concept with a help of other concepts. This sets the domain for the concept of interest. Like other researchers we consider user satisfaction to be an attitude. Following Melone (1990) who adapted the definition of user attitude to the area of information systems, we define user attitude in our context as a predisposition to respond favorably or unfavorably to a web site. We further adopt the assumption of the theory of reasoned action according to which attitudes influence behavior or at least behavioral intention (Ajzen/Fishbein 1980). This framework does not negate, in our view, the reciprocal influence of behavior on attitude or the possibility that user attitudes are not always accessible to the users as suggested by the cognitive view of user attitudes. The same is true for the possibility that user attitudes sometimes also serve nonevaluative purposes. This is the case, for example, when users express attitudes that are assumed to be shared by a peer group independently of the attitude object. In this work we try to establish a proven way to measure USW as a first step. Some measurements of behavior that might influence the attitude toward a web site have also been conducted as described in section 6.

An operational definition specifies the rules by which the concept is measured. Our operational definition is: Satisfaction of a visitor of a web site with the site is measured through a weighted sum of visitor’s attitudes to 17 items. The weight of an item is the importance assigned to it by a visitor. The corresponding formula is:

$$USW = \frac{\sum_i x_i w_i}{\sum_i w_i} \qquad \text{where}$$

x_i = evaluation of item i by a visitor

w_i = importance of item i assigned by a visitor

The satisfaction of a group of evaluators with a site is measured by the average of individual USW scores.

Like with any information system *ease of use* is important for a web site since it enables visitors to make use of it. Otherwise users will quickly leave that site. Within a company users often must use the systems they are offered. In the world-wide web users are like consumers who have a wealth of services to choose from. Therefore, they must be repeatedly attracted to the web site by being offered rich *information content* added by more or less *entertainment*. Since one of the main characteristics of the web, as opposed to mass media like TV, radio, or newspapers, is its capability of two-way communication users have come to expect *interactivity* from web sites. We assumed that these four factors lead to USW and identified items to represent them based on the literature on user information satisfaction (e.g., Bailey/Pearson 1983) and existing Internet technologies and services. Users evaluate each item on a 7-points Likert scale.

In the rich literature on user information satisfaction (UIS) the surveyed users were in most cases part of the organization that owned and often developed the information system(s) under evaluation. Therefore, many questions in these instruments relate to user participation in the development process, their relationship with the MIS department, or the training they received in order to work with the observed systems. In USW such questions do not make sense in the general case (in particular cases, such as in intranets, they can be relevant). The instrument would also need to be adapted for use in established business-to-business relationships where parts of a web site are made available only to selected business partners over the Internet or an extranet.

The instrument in its original form as used in the pretest is given in fig. 1. Later amendments to it are described where the reasons for the change are explained. In an evaluation cycle in which the same user rates several sites one extra questionnaire is given to determine the importance of items. The only difference compared to the instrument in fig. 1 is that the end points of the Likert scale are named unimportant and very important. If only one site is being evaluated then the importance question is integrated with item evaluations.

4 Instrument Validation

4.1 Reliability

The reliability of an instrument is most often measured by the Cronbach-α statistic. This statistic tries to determine whether the items really measure the same construct. If they do one rater's ratings of these items should be highly correlated. This is also true for groups of items that constitute a factor. The factors can be derived through a factor analysis as done in section 4.3. For the pretest instrument a Cronbach-α of 0.8644 was computed. This already satisfactory value was raised to 0.8938 after the below described changes to the instrument.

The validity of an instrument can be examined along different dimensions. These are content, construct, and criterion-related validity.

4.2 Content Validity

Content validity relates to the question whether the construct really measures what it intends to. Such a claim can be backed up by theory and pretesting. Pretesting was performed with a help of student subjects. The students of a course on ”Introduction to MIS” were asked to visit and rate ten different sites. They all received basic instructions into the use of the web and other standard Internet technologies. In order to guarantee a minimal exposure to a site they were asked to perform a specific task at each site. Such tasks were, e.g., the search for a specific document in the online archives of a news magazine or the configuration of a desired car on the site of a car manufacturer. The quantitative and qualitative analyses of the pretest results and the subsequent modifications of the instrument can be considered a proof of content validity.

Ease of use			
1. Response speed	Very bad	○○○○○○○	Very good
2. Navigation support	Very bad	○○○○○○○	Very good
3. Choice between graphics and text	Very bad	○○○○○○○	Very good
4. Choice between presentation with or without frames	Very bad	○○○○○○○	Very good
5. Use of new web technologies (e.g., Java, virtual reality)	Very bad	○○○○○○○	Very good
Information content			
1. Quantity	Very bad	○○○○○○○	Very good
2. Quality	Very bad	○○○○○○○	Very good
3. Currency	Very bad	○○○○○○○	Very good
4. Integration with other communication of the company (e.g., TV spots)	Very bad	○○○○○○○	Very good
5. Links to other web sites with or without relationship to company products (e.g., product tests or cultural events)	Very bad	○○○○○○○	Very good
6. Database queries	Very bad	○○○○○○○	Very good
7. Customized information (e.g., the calculation of an individual insurance policy)	Very bad	○○○○○○○	Very good
8. Availability of downloadable documents, programs, forms	Very bad	○○○○○○○	Very good
Entertainment value			
1. Amusing	Very bad	○○○○○○○	Very good
2. Exciting	Very bad	○○○○○○○	Very good
Interactivity			
1. Transaction support	Very bad	○○○○○○○	Very good
2. Active requests for e-mail contacts	Very bad	○○○○○○○	Very good
3. Dialogue possibility via live-chats (synchronous communication)	Very bad	○○○○○○○	Very good
4. Dialogue possibility via black-boards or asynchronous discussions groups	Very bad	○○○○○○○	Very good
Overall satisfaction	Very bad	○○○○○○○	Very good

Figure 1: Original USW Survey Instrument

4.3 Construct Validity

Construct validity refers to the extent to which the instrument measures a theoretical construct. It is most often examined via factor analysis as recommended in literature on behavioral research (e.g., Kerlinger 1978). A factor analysis of the data from the pretest revealed a structure of four factors that explained 55.5% of the total variance. However, the assignment of items to factors was slightly different than we expected it. Table 2 shows the factors, the names we gave them, and items with factor loadings over 0.5. The exact loadings are given in parentheses. The extraction technique used was principal components analysis and the rotation method applied was varimax. The table also shows the values of Cronbach-α for each factor. Based on recommendations by Nunnally (1978) these values can be considered acceptable.

Factor	Ease of use	Information content	Additional information	Interactivity
Cronbach-α	0.6886	0.8202	0.7550	0.8016
Items	Speed (0.65)	Quality (0.77)	Databases (0.70)	Live-chats (0.85)
	Graphics (0.65)	Exciting (0.75)	Downloads (0.63)	Discussion groups (0.84)
	Navigation (0.63)	Currency (0.72)	Customization (0.62)	E-mail (0.66)
	Frames (0.53)	Amusing (0.71)	Integration (0.55)	Transactions (0.53)
		Quantity (0.69)	Frames (0.55)	
			Links (0.52)	

Table 2: Factors of User Satisfaction with Web Sites

While the factor analysis indicated that the instrument is already acceptable we wanted to improve it based on the results of the reliability and factor analyses. A decrease of the number of items seemed to be also desirable for practical reasons (as some raters complained about the length of the instrument). We decided to drop items that had a low multiple r^2 with other items (below 0.3) *and* a low factor loading (below 0.6) from the instrument in its subsequent use. This was the case with two items: ”use of new web technologies” (that did not load on any factor) and ”integration with other communication of the company”. The item ”choice of presentation with or without frames” loaded on two factors though in both cases with a loading factor in the low fifties. Since all widely used browsers meanwhile handle frames without problems it does not make much sense to use it further. We exchanged this item for a new one: ”keywords search within the site”. In later uses of the instrument this new item loaded on the factor ease of use (factor loading 0,65), its multiple r^2 with other items was acceptable (0.4346). These amendments led to an instrument with 17 items. Other items are only omitted when their use does not make sense because the observed site does not offer the relating feature (e.g., live chatting).

Factor analyses in later uses of the instrument always revealed an underlying structure of four factors. This is also true when the grouping of items presented to the raters was different than in fig. 1 with respect to number of groups and their sequence. The amount of variance explained and the reliability values of the individual factors were somewhat higher. However, the assignments of items to factors varied slightly.

4.4 Criterion-related Validity

Criterion-related validity can be studied in different ways, e.g., as predictive or correlative validity. An instrument shows predictive validity if it can be used to predict a future event. Correlative validity is given if the resulting values highly and significantly correlate with values derived from measurements of the same phenomenon but with a different construct. In order to test criterion-related validity the raters in the pretest were asked to name the site they liked the best among the ten sites they evaluated with our instrument. This choice was made on a piece of paper, i.e., independently of the site evaluations that were done online. Then the sites were ranked by the number of top spots. These ranks were correlated with ranks obtained on the basis of average USW scores for each site. The rank correlation coefficient was 0.8504, significant at $p<0.01$. Given this fairly high and significant value correlative validity can be considered to have been established.

In the measurement of user information satisfaction correlative validity has often been examined by asking the rater after rating individual items to give his overall satisfaction with the rated system characteristic (e.g., Bailey/Pearson 1983) or the whole information system. This value has then been correlated with the measure calculated based on the rating of individual items. This approach has been applied here too. The correlation was highly significant ($p<0.01$) with a coefficient of 0.6031. However, this test is weak because it does not really fulfill the demand that the same phenomenon be measured with a different construct. Since the stronger test described above could not always be administered (e.g., when only one site was being measured) this simple test was performed as a second best solution in the subsequent applications of the instrument. In all cases there was a significant correlation between this rating and our measure of USW with a coefficient of $r=0.75$ or higher.

Predictive validity has not been examined as often as correlative validity in the work on UIS. Sometimes the UIS score is correlated with the (perceived) use of the observed information system. A similar test would make sense here as well, e.g., the correlation of visits to a site over a given period of time with the users' USW score. Unfortunately, the data for such a test are not easy to obtain. On sites without registration requirement it is not possible to identify individual visitors over time. In Germany, where the instrument was developed and used, even sites that require registration (e.g., some online services within the Internet) do not track individual users, at least officially, as this is severely restricted by data privacy laws. Individual customer data may only be used for accounting purposes but may not be used for profiling or other types of tracking and may not be stored for longer time periods than it is necessary for the original purpose of data collection.

5 Measurement Results

Two significantly different measurement setups were used. In the first setup student subjects surf the web sites at least as much as to accomplish a given task. Then, they fill out the evaluation form on our server. We refer to this setup as the "offsite" measurement. It can obviously be done without the cooperation of the evaluated site. The advantage of this setup is that each rater has the same minimum exposure to the site, the disadvantage is that all raters are students and that they are "forced" to perform the evaluations (incomplete ratings were sufficient to pass the exercise but they were disregarded in further analyses if less than 50% of items were evaluated). In the second setup, we refer to as "onsite" evaluation, the evaluation form is placed within the web site to be evaluated. The completed evaluations are immediately and automatically sent to us by e-mail. The advantage of this setup is that answers can be received from a sample of the entire Internet population. The disadvantage is that the raters usually have very different histories of visits at the site. This can be controlled if the visitors to the site can be identified over time (e.g., via cookies). Otherwise one needs to rely on usually unreliable statements about the perceived use of the site. While the majority of sites were evaluated offsite more than 50% of the raters were drawn from onsite evaluations, i.e., they belong to the general Internet population.

Although the goal of our work was the development of the described instrument it may be of interest to see some of the results of our measurements. Table 3 identifies the evaluated sites except in two cases where measurements were done onsite, it shows the average ratings of the sites and the standard deviations, it shows whether the difference in evaluation scores between two consecutive sites on the list is significant for sites that can be considered competitors, and it gives some additional information about the evaluations. The number of raters varies even when evaluations are performed in the same cycle because incomplete evaluations were disregarded as explained above.

Industry	Company or Product	Rating Mean	SD	Difference significant ?	Number of Raters	Evaluated in
Media (Magazine)	Spiegel	4.563	0.809	No	120	May 1997
Media (Magazine)	Focus	4.487	0.729		122	May 1997
Media (TV)	ZDF	4.565	0.796	Yes	159	May 1998
Media (TV)	ARD	4.362	0.797	No	160	May 1998
Media (TV)	Pro7	4.362	0.774	Yes	161	May 1998
Media (TV)	SAT1	4.140	0.928		160	May 1998
Finance (bank)	Bank 24	4.296	0.836	No	122	May 1997
Finance (bank)	Advance Bank	4.189	0.879		119	May 1997
Retail	My-world (Karstadt)	4.396	0.979	No	122	May 1997
Retail	Quelle	4.315	0.844	Yes	122	May 1997
Retail	Otto	3.978	0.831	No	177	Nov. 1997
Retail	Neckermann	3.943	0.777		183	Nov. 1997
Food (ice cream)	Langnese	4.481	0.812	Yes	175	Nov. 1997
Food (soups)	Maggi	4.316	0.890	Yes	172	Nov. 1997
Food (coffee)	Jacobs	4.182	0.952	Yes	178	Nov. 1997
Food (chocolate)	Ritter Sport	3.678	1.010		159	May 1998
Restaurants	Mc Donald’s	4.043	0.986		160	May 1998
Packaged goods	Henkel	4.111	0.902	Yes	122	May 1997
Packaged goods	Procter & Gamble	3.911	1.030		120	May 1997
Manufacturing	Mercedes Benz	4.431	0.969	Yes	122	May 1997
Manufacturing	Audi	4.186	0.934		121	May 1997
Media (on-line service)		5.08	0.894		31	Dec. 1997
Media		4.8	0.800		561	Sep. 1997

(radio)							

Table 3: Results of Evaluations with the USW Instrument

The two onsite measurements revealed the highest USW scores. One reason for this result could be that surfers who are not satisfied with the site do not visit it anymore and therefore the results are biased by the sample composition. Since respondents were offered prizes in these two cases it could also be that respondents (wrongly) assumed that too critical evaluations would decrease their chances of winning a prize. The number of onsite measurements is too small, of course, to enable us to make final judgements. The low number of participants in the evaluation of the online service is due to the facts that at that time the service only had about 2,000 subscribers and that the prize offered was a book. In the case of the radio station the prize was 500 German Mark in cash.

It is also interesting to see where the discrepancies between the importance visitors assign to an item and their assessment of its realization occur. For an individual site such a comparison can help to identify areas for improvement. In fig. 2 the average scores for the evaluation cycle in May 1998 are given. The results are representative of other evaluations we performed, including those done onsite. The items are grouped by factors given in table 2, the differences with respect to items shown reflect the discussed improvements of the instrument.

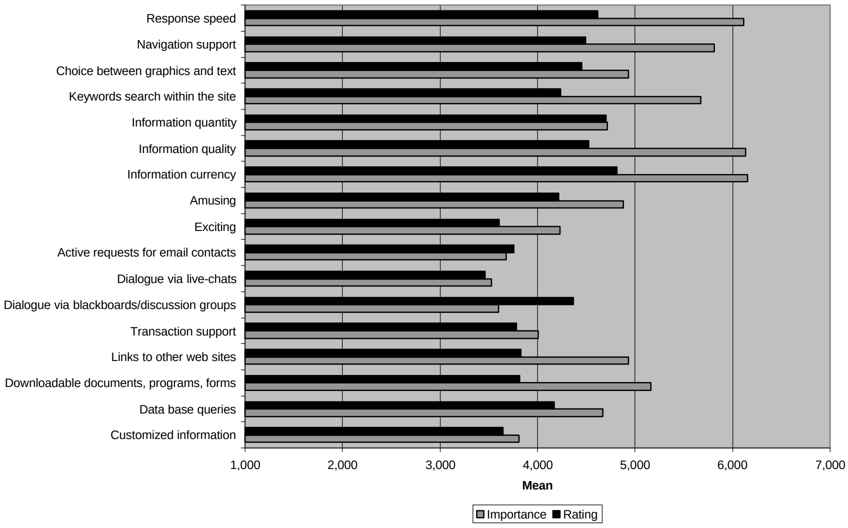

Figure 2: Importance versus Rating of Items

Web users seem to be most concerned themselves with information content and ease of use of a web site. The biggest gaps (above 1) exist with respect to information quality, response speed, keywords search, availability of downloadable information, information currency, navigation support, and links to other sites in that order. The response speed depends to a big extent on the equipment of the user and his connection to the Internet; it is, therefore, only partly controllable by the site owners. However, web sites should play their part in making the most efficient use of the available Internet resources and technologies. It appears that there is a bit of a paradox with respect to information quantity. On one hand users feel that they do not need more information. On the other hand they would like to have more downloadable information and more links to other web sites. The solution of this paradox is probably that users want more ”good” information without increasing the amount of information they are already receiving. The demands for dialogue and discussions are already adequately met. These items also do not rank high with respect to their importance. This does not necessarily mean that Internet users are not eager to communicate with each other, but that they may not be so keen on it in the context of commercial web sites.

6 Correlates

Variances in evaluations may exist not only with respect to sites but also with respect to specific groups of raters. Therefore, in each evaluation (cycle) additional questions about the raters were asked in order to analyze correlates of USW. At this point in time it is too early to make assumptions about possible directions of influence and more elaborated modeling. Our findings can be a basis for formulating hypotheses that can be examined through the further use of the instrument.

Demographics

A significant difference in evaluations by men and women occurred in only four of 23 cases. None of the sites evaluated was specifically targeted to either of the sexes. In all four cases where a significant difference existed women evaluated the sites more favorably. These sites were Jacobs, Langnese, Otto, and the online service.

Since all offsite evaluations were done by students the possibility exists that other occupational groups would evaluate sites differently. We used the two onsite evaluations to address this question. No significant difference between students' and other groups' evaluations could be determined! This indicates that the instrument can be used with the general Internet population although it has been originally developed based on student evaluations.

Internet experience

One could expect that more experienced users would have higher expectations than the beginners and would therefore be more critical than the beginners. Our findings were, however, that where a significant difference existed more experienced users (more than six months of Internet use at the time of evaluation) were giving slightly better scores. This could be due to the fact that they can find the desired information easier than the beginners.

In onsite evaluations we also asked how many pages on that site the surfers visit usually. Based on this information we examined whether the ”deep surfers” are more satisfied than the ”surface surfers”. This was the case though the difference in rating was not high. The same relationship holds for the comparison between those users who have visited the site many times before the rating and those users who are relatively new to the site. The raters that put the examined site on their list of bookmarks or favorites were not significantly more satisfied with it. The explanation could be that the addition of a site to the list happens after one of the early visits to a site rather than after extensive exposure to it. Many bookmarks may be seldom used. The way users learned about the URL of a site was a source of significant differences in USW scores. Users who heard the URL on radio (in the case of the radio station) or who received it from an acquaintance were more satisfied with the site.

Observations in this subsection can also be interpreted with the view that behavior influences attitudes. Then, based on our preliminary findings it can by hypothesized that more experience with web sites in general and more exposure to a site lead to higher USW.

Habits

One question that especially intrigued us was whether product preferences automatically translate to web site preferences. Thus, we asked the raters in one evaluation cycle about their favorite TV network and about their consumption of hamburgers and chocolate. In the latter case raters who consumed more hamburgers or more chocolate than the other raters were not significantly more satisfied with the web sites of McDonald’s or Ritter Sport. In the case of TV networks or stations the situation was not so clear. Those surfers who watched TV more than the other raters gave overall higher USW scores to web sites of TV networks or stations. However, the preference for a specific TV station was in only one of four cases (ZDF) correlated with significantly higher USW scores for this station’s web site. Interestingly, this station was mentioned the least as the preferred TV station.

These observations lead us to postulate that product preferences do not automatically translate into web site preferences. On the one hand, if a firm with popular products wants its customers to keep visiting its web site it is not enough just to be present on the web. On the other hand, web sites can probably ”live” a life on their own and attract their followers independently of how popular the firm that operates it may be outside of the web.

7 Outlook

The instrument developed seems to be valid and reliable. Still there is room for improvement and further testing of its psychometric qualities. The latter is, for example, the case with respect to predictive validity. The hypotheses formulated on the basis of the reported measurements need to be further examined. As indicated in the brief discussion of the theoretical aspects of user attitudes other views of their structure exist that are worth exploring. To that extent further measurements of user behavior and its impact on USW are necessary. Further, it should be examined under which circumstances USW has an actual impact on user choices of a web site. Finally, more complex models can be developed when valid and reliable measures of variables like web site use or web site effectiveness become available. These models could then be used to understand the complete process of value creation on the web.

References

Ajzen, I./Fishbein, M. (1980): Understanding Attitudes and Predicting Social Behavior. Prentice-Hall, Englewood Cliffs, NJ, 1980.

Alpar, P. (1997): Nutzung des Internet bei deutschen Unternehmen. Nachrichten für Dokumentation (NfD), 48 (1997)3 Mai-Juni, 180–185.

Bailey, J. E./Pearson, S. W. (1983): Development of a Tool for Measuring and Analyzing Computer User Satisfaction. Management Science, 29 (1983)5, 1983, 530-545.

Boling, E. (1995): Usability Testing for Web Sites. Presented at the 7th Annual Hypermedia ’95 Conference, downloaded on 9/8/98 from http://www. indiana.edu/~iirg/ARTICLES/usability/usability.main.htm.

Bonanno, D. M. (1996): Evaluating Electronic Communication Patterns Over a Semester: A Qualitative Content Analysis Downloaded on 09/03/98 from http://trochim.human.cornell.edu/webeval/webcomm/webcomm.htm.

Bowen, K. B. (1996): Website Evaluation: Experimental and Quasi-Experimental Design Issues. Downloaded on 9/3/98 from http://trochim.human.cornell. edu/webeval/webexper/webexper.htm.

Fitzelle Jr., G. T./Trochim, W. M. K. (1996): Survey Evaluation of Web Site Instructional Technology: Does it Increase Student Learning? Downloaded on 9/3/98 from http://trochim.human.cornell.edu/webeval/webques/ webques.htm.

Ho, J. (1997): Evaluating the World Wide Web: A Global Study of Commercial Sites. J. of Computer-Mediated Communication, 1 (1997)3 June, downloaded on 09/03/98 from http://www.usc.edu/dept/annenberg/vol3/issue1/ho.html.

Jarvenpaa, S. L./Todd, P. A. (1997): Consumer reactions to electronic shopping on the World Wide Web. Int. J. of Electronic Commerce 1 (1997)2, 59-88.

Jaspersen, T. (1996): Erfolgskontrolle von Online-Marketing-Aktivitäten – ”Hit-Liste”. Absatzwirtschaft 12, 1996, 46-48.

Kerlinger, F. N. (1973): Foundations of Behavioral Research. New York 1973.

Lohse, G. L./Spiller, P. (1998): Quantifying the effect of user interface design features on cyberstore traffic and sales. In CHI’98 Conference Proceedings. ACM Press, Los Alamitos, Cal.

Melone, N. P. (1990): A Theoretical Assessment of the User-Satisfaction Construct in Information Systems Research. Management Science, 36 (1990)1, 76-91.

No author (1998): WA – Findings. Downloaded on 09/03/98 from http://www.businessmedia.org/businessmedia.nsf/pages/wa_results.html.

Nunnally, J. (1978): Psychometric Methods. 2nd ed. McGraw-Hill, New York 1978.

Selz D./Schubert, P. (1997): Web Assessment – A Model for the Evaluation and the Assessment of successful Electronic Commerce Applications. Int. J. of Electronic Markets, 7 (1997)3, 46-48 and accompanying paper downloaded on 05/27/98 from http://www.electronicmarkets.com/em97_3wa.html; a version also appeared in Proceedings of the 31st HICSS Conference, Hawaii, 1998.

Trochim, W. M. K. (1996): Evaluating Websites. Downloaded on 09/03/98 from http://trochim.human.cornell.edu/webeval/webintro/webintro.htm.

Trochim, W. M. K./Hover, D. (1996): Mapping Student Views of the Benefits of a Course Website. Downloaded on 09/03/98 from http://trochim.human.cornell.edu/webeval/mapuser/mapuser.htm.