Assessing a Firm's Web Presence: A Heuristic Evaluation Procedure for the Measurement of Usability

Ritu Agarwal • Viswanath Venkatesh
Robert H. Smith School of Business, University of Maryland, Van Munching Hall, College Park, Maryland 20742-1815
ragarwal@rhsmith.umd.edu • vvenkate@rhsmith.umd.edu

Web site usability is a critical metric for assessing the quality of a firm's Web presence. A measure of usability must not only provide a global rating for a specific Web site, ideally it should also illuminate specific strengths and weaknesses associated with site design. In this paper, we describe a heuristic evaluation procedure for examining the usability of Web sites. The procedure utilizes a comprehensive set of usability guidelines developed by Microsoft.

We present the categories and subcategories comprising these guidelines, and discuss the development of an instrument that operationalizes the measurement of usability. The proposed instrument was tested in a heuristic evaluation study where 1,475 users rated multiple Web sites from four different industry sectors: airlines, online bookstores, automobile manufacturers, and car rental agencies. To enhance the external validity of the study, users were asked to assume the role of a consumer or an investor when assessing usability. Empirical results suggest that the evaluation procedure, the instrument, as well as the usability metric exhibit good properties. Implications of the findings for researchers, for Web site designers, and for heuristic evaluation methods in usability testing are offered.
(*Usability*; *Heuristic Evaluation*; *Microsoft Usability Guidelines*; *Human-Computer Interaction*; *Web Interface*)

"On the Web, users experience usability first and pay later." (Nielsen 2000, p. 11)

1. Introduction

In an economy witnessing explosive growth in consumer electronic commerce (Hoffman and Novak 2000) and net-enabled organizations (Straub and Watson 2001), it is no surprise that Web site design represents an issue of considerable importance to firms. An increasing number of businesses are choosing the Web as an alternative channel for developing a brand reputation, for transacting with and servicing customers and investors, or simply for public relations purposes (Subramaniam et al. 2000). Therefore, significant managerial attention is being focused on the experience that consumers have in cyberspace when they visit a corporate Web site. Although the Web, by virtue of its multimedia capabilities, provides an opportunity for a firm to offer a unique and satisfying experience to its Web site visitors (Hoffman and Novak 1996), developing a corporate Web site is not without risks. In particular, the design of the Web site is a crucial determinant of whether visitors are likely to return to the site (Klein 1998) and, indeed, of consumer satisfaction with Internet shopping. A critical challenge facing businesses today, then, is to develop a Web presence that is not only compelling for the visitor, but is also able to serve his or her instrumental goals well.

Information Systems Research, © 2002 INFORMS
Vol. 13, No. 2, June 2002, pp. 168–186
1047-7047/02/1302/0168$05.00
1526-5536 electronic ISSN

How can an organization measure the quality of its Web presence? More significantly, what is an appropriate metric that not only evaluates Web site quality but also provides managers with insights into potential problem areas? To answer these questions, we examine Web sites through a human-computer interaction (HCI) lens. Specifically, we focus on a key concept that emerges from HCI research—that of usability. Although the notion of usability, has been defined in a variety of ways by scholars (see Nielsen 1994, Gray and Salzman 1998), prior research overwhelmingly suggests that usability is associated with many positive outcomes, such as a reduction in the number of errors, enhanced accuracy, more positive attitudes toward the target system, and increased usage (Lecerof and Paterno 1998, Nielsen 2000). Therefore, we argue that usability is likely to be a key and proximal metric for evaluating the success of an organization's Web presence. A compelling Web presence, in turn, should contribute to the short- and long-term success of Web sites by encouraging repeat visitors and contributing to customer satisfaction (Klein 1998, Lam and Lee 1999). Indeed, empirical work by Lohse and Spiller (1999) shows that interface features, such as those assessed during usability testing, explain substantial variance (61%) in sales for online stores.

A procedure and an accompanying metric for assessing Web site usability must exhibit several important properties. Not only should the metric be able to discriminate across sites that exhibit varying levels of usability, it must also offer specific insights into areas of weaknesses in the design of the site. Ultimately, the goal for a useful, good usability metric is to help improve an organization's Web presence. It is particularly important, therefore, for a metric to provide detailed information about a company's Web presence such that the Web presence can be benchmarked against competitors' Web sites to understand relative strengths and weaknesses. Then, based on the usability assessment of the focal site, designers can draw upon specific design principles, such as those offered by Nielsen (2000) to improve the Web site.

The purpose of this paper is to describe a method for assessing Web site usability and an accompanying metric that is likely to be of value to both researchers and practitioners. We present: (1) a detailed discussion of an evaluation procedure (Microsoft Usability Guidelines) to assess usability, (2) a method to apply this procedure in practice, (3) details of the instrument development process, and (4) an extensive field application of the method and the instrument to establish external validity. The field application was conducted using 1,475 users' assessments of multiple Web sites from four different industry sectors—i.e., airlines, online bookstores, automobile manufacturers, and car rental agencies.

1.1. Human-Computer Interaction and Usability

As observed earlier, the notion of usability is a key theme in the HCI literature. Research in the HCI tradition has long asserted that the study of human factors is key to the successful design and implementation of technological devices (e.g., Shneiderman 1980 and 1998). The overarching goal of a majority of the HCI work has been to propose techniques, methods, and guidelines for designing better and more "usable" artifacts. To this end, researchers have examined diverse phenomena such as the design of programming languages (Sime et al. 1973), errors made while utilizing alternative modeling approaches (Agarwal et al. 1999), and the usability of operating systems such as UNIX (Jeffries et al. 1991). Drawing upon cognitive frameworks of human-computer interaction grounded in psychology, prior research developed user models that delineate the cognitive structures driving user behavior (Card et al. 1983). Researchers also focused attention on explicating how users coordinate knowledge between the task domain and the device domain (Payne et al. 1990). Two important findings from this work are: (1) the importance of consistency in design and (2) the idea that prior knowledge possessed by users plays a key role in subsequent learning of new artifacts and devices.

Usability has been conceptually defined and operationally measured in multiple ways. Gray and Salzman (1998, p. 238) succinctly summarize the state of affairs related to the definition of usability noting that "the most important issue facing usability researchers and practitioners alike [is] the construct of usability itself." Definitions of usability range from the high-level conceptualization incorporated in the ISO 9241 standard (Karat 1997) to more focused descriptions that include

notions of user relevance, efficiency, user attitude, learnability, and safety (Lecerof and Paterno 1998). In detailing their concept of usability, Lecerof and Paterno underscore that the most critical aspect of usability is contingent upon the actual system. For example, ease of use might be a primary criterion for systems designed for use by children, while efficiency is likely to be a major usability goal in the design of banking systems. For the purposes of our research, we adopt the ISO definition of usability—"the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (Karat 1997, p. 34).

2. Literature Review

A variety of alternative approaches to usability evaluation have been proposed in prior work. Nielsen (1994) identify eight distinct approaches: heuristic evaluation, guideline reviews, pluralistic walkthroughs, consistency inspections, standards inspections, cognitive walkthroughs, formal usability inspections, and feature inspections. In an alternative taxonomy, Gray and Salzman (1998) classify usability evaluation methods into analytic and empirical categories, where the former includes approaches such as heuristic evaluation (Nielsen 1994), cognitive walkthroughs (Polson et al. 1992), guidelines (Blatt and Knutson 1994), and GOMS (Card et al. 1983), whereas the latter category refers to all methods generally termed as "user testing."

There are two recurrent themes in all of these approaches to usability evaluation. One is the notion that usability is multifaceted and must be assessed by using a variety of different measures. A second common characteristic of usability evaluation methods is their dependence on subjective assessments in the form of user judgments. Thus, usability is not intrinsically objective in nature, but rather is closely intertwined with an evaluator's personal interpretation of the artifact and his or her interaction with it. Nonetheless, all usability evaluation approaches begin with the basic assumption that it is possible to identify, at varying levels of granularity, what the features of a "usable" system might be. For example, in the context of Web site design specifically, Nielsen (2000) offers a wide range of design principles for a usable system, derived from a synthesis of extensive prior work, which he and his colleagues conducted. In other work, Keevil (1998), based on a review of existing usability guidelines, developed a checklist of Web site features that evaluators could respond to in a "yes or no" format. Responses to the checklist are then used to compute an overall usability index for the Web site.

With regard to evaluation methods used for Web site usability assessments, Kantner and Rosenbaum (1997) observe that heuristic evaluation and laboratory testing are two of the most frequently used approaches. Heuristic evaluations are assessments conducted by a small group of evaluators against a preestablished set of guidelines or "heuristics" (Nielsen 1994). The evaluators are generally experts in usability, although it is desirable to use individuals who are both usability and domain experts (Kantner and Rosenbaum 1997). In contrast, laboratory testing utilizes real users as subjects and provides detailed insight into specific problems and issues that users face while interacting with the target Web site.

Researchers have proposed many different dimensions along which Web sites could potentially be evaluated. For instance, Eighmey and McCord (1998) examined audience experience of Web sites on 80 evaluative statements across 5 Web sites spanning a range of products and industries. Seventeen factors arising from the results were subsequently reduced to nine groups, including personal involvement, useful information, simplicity of organization, and desire for relationship. In other work, Gehrke and Turban (1999) identified five major categories of factors that ought to be considered while designing Web sites for business: page loading, content, navigation efficiency, security, and a consumer/marketing focus. Katerattanakul and Siau (1999) focused on one specific aspect of Web site design—i.e., information quality—and proposed a framework comprised of four information quality categories: intrinsic, contextual, representational, and accessibility information quality. Finally, Chau et al. (2000) examined the effects of different modes of information presentation on the use of online shopping, arguing that the relative importance of graphics versus

text in the design of a Web site is likely to vary with user familiarity with product items.

In summary, the literature points to the importance of usability in the design of user interfaces and identifies several dimensions along which Web sites can be evaluated for usability. While each approach to usability assessment adopts a unique perspective on the phenomenon and has its strengths and limitations, one of the key limitations that we observed was the lack of close ties to actual design practice by market leaders. Finally, we found that extensive field tests that examined multiple user roles, multiple industries, and a large sample of users were almost completely absent.

2.1. Microsoft Usability Guidelines

In our work, we employ the set of heuristic guidelines of a market leader, the Microsoft Usability Guidelines (MUG). A white paper by Keeker (1997) describes the various guidelines and also outlines some additional key information about MUG. Although we study a particular set of dimensions to assess usability (i.e., those prescribed in MUG), the general methodology presented here is robust enough to be used with alternative evaluation criteria. Providing a comprehensive basis for the heuristic evaluation of Web sites, the Microsoft Usability Guidelines are organized around five major categories: content, ease of use, promotion, made-for-the-medium, and emotion. These categories are expected to cover the range of usability-related aspects of a Web site. Additionally, a close study of the guidelines reveals that four of the five categories have subcategories that are meant to represent various dimensions of the major category (see Keeker 1997), thus providing greater assurance that the content or domain of the construct (i.e., usability) is adequately covered by these dimensions. While we will discuss the detailed instrument development later in this section, we now present the conceptual definitions of various categories and subcategories.

2.2. Categories and Subcategories

The five major categories and their definitions are:

Content assesses the informational and transactional capabilities of a Web site. This category is closest to constructs such as perceived usefulness (Davis et al. 1989, Venkatesh and Davis 2000) and relative advantage (Agarwal and Prasad 1997, Moore and Benbasat 1991) examined by researchers as antecedents to various technology acceptance outcomes.

Content comprises four subcategories. Although perceived usefulness and relative advantage (i.e., the constructs most similar to content) have been conceptualized and operationalized as unidimensional, MUG suggests multiple subcategories that, in fact, capture various aspects associated with content. These subcategories are: (1) *relevance,* relating to the pertinence of the content to the core audience; (2) *media use,* signifying the appropriate use of multimedia content; (3) *depth and breadth*, examining the appropriate range and detail of topics; and (4) *current and timely information*, capturing the extent to which a Web site's content is current.

Ease of use relates to the cognitive effort required in using a Web site. The construct of ease of use has been employed extensively in IT research (Davis et al. 1989, Venkatesh 2000) and, together with perceived usefulness, has been shown to be an important predictor of technology acceptance outcomes. In MUG, ease of use comprises three subcategories. As with perceived usefulness, prior research conceptualized ease of use as unidimensional, but MUG conceptualizes the following three subcategories of ease of use: (1) *goals*, relating to clear and understandable objectives; (2) *structure*, focusing on the organization of the site; and (3) *feedback*, capturing the extent to which the Web site provides information regarding progress to the user.

Promotion captures the advertising of a Web site on the Internet and other media. Although not a direct outcome of design decisions made regarding a specific Web site, promotion is critical to drive traffic to the site. In MUG, promotion is not broken down into subcategories.

The fourth category, *made-for-the-medium* relates to tailoring a Web site to fit a particular user's needs. The Web offers unprecedented opportunities for mass customization, and personalization is a critical requirement of Web sites today. Indeed, contemporary marketing strategies such as relationship (Day 2000) and one-to-one marketing (Peppers and Rogers 1999) require that Web sites not be static in design; rather, they should provide dynamic content that is tailored to the unique and idiosyncratic needs of a specific user.

Made-for-the-medium has three subcategories: (1)

community, capturing if the Web site provides users with an opportunity to be part of online group; (2) *personalization*, reflecting the technology-oriented customization of the Web site; and (3) *refinement*, relating to the particular prominence given to current trends.

Finally, *emotion* taps into affective reactions invoked by a Web site. Affective responses have been shown to play an important role in computer use situations (Agarwal and Karahanna 2000, Venkatesh 2000, Venkatesh and Speier 2000, Webster and Ho 1997). MUG views four subcategories as being the components of emotion: (1) *challenge* captures the idea of difficulty, particularly as it relates to a sense of accomplishment, rather than simply functional complexity or obscurity; (2) *plot* relates to how the site piques the user's interest, especially with a story line; (3) *character strength* relates to the credibility conveyed by the site, particularly via the individuals portrayed on the site; and (4) *pace* examines the extent to which the site provides users an opportunity to control the flow of information.

Web sites present a distinct challenge when it comes to usability assessment. Unlike most other software, which has a reasonably well-defined audience with a limited set of tasks that a user can perform, visitors arrive at a Web site for a multitude of reasons. In an increasingly Web-centric consumer environment, a visitor to a Web site may play the role of an information seeker, a "surfer," or a serious consumer desirous of transacting (Breitenbach and Van Doren 1998, Olsson 2000). As noted by Kantner and Rosenbaum (1997, p. 153), "the definition of user . . . is becoming more vague [sic] because anyone can access the site." Each user role and the associated user goals (e.g., information seeker, surfer, transactor) embody a unique set of requirements and needs with regards to the design of the site.

Besides challenges arising from heterogeneity of the user population, Web sites have other distinct characteristics when compared with traditional software-user interfaces that further introduce complexities in design and usability testing (Nielsen 2000, Shneiderman 1998). For instance, there is considerable diversity in the devices through which users can access Web sites, ranging from a cellular telephone to a television set to a personal digital assistant. Design decisions as well as usability testing need to take such device diversity into account. Designers have limited control over user interaction in that users can typically choose whatever path they like to navigate through pages. This is in contrast to traditional software where certain options may be rendered unavailable at different times during the user's interaction with the software (such as the graying out certain menu options.) It is, therefore, not surprising that assessing the usability of Web sites is not simply an application of software evaluation methods, but rather one that requires a new perspective.

Based on our review of the literature and the discussion above, we conclude that MUG provides a comprehensive range of categories and subcategories that allow users to clearly discriminate across industries and products. Further, the range and depth of detail covered by MUG provides a basis for a user to discriminate across Web sites *within* a particular industry. Together, this facilitates the generation of information necessary to compare Web sites both across industries and within an industry. Finally, the breadth of the guidelines will help identify specific areas of strengths and weaknesses in the design of the target Web site.

3. Study Methods and Procedures

MUG prescribes a set of evaluative criteria that help assess Web site usability. However, as argued earlier and suggested by the conceptual definition of usability, not all criteria are likely to be equally important across different types of users and Web sites. Thus, it is important that a usability assessment procedure provides detailed information on what matters to different types of users when they visit Web sites from different industries. The relative importance of such evaluative criteria can be established through direct methods (such as the constant sum scale that asks consumers to allocate a fixed number of points, usually 100, to his or her evaluative criteria as an indicator of their importance) or through indirect methods such as conjoint analysis (Hawkins et al. 1995). We developed a method for the assessment of usability that includes weights and ratings. First, an "evaluator" (i.e., Web site user) provides the relative importance (weights) of the different categories. In this step, evaluators distribute 100 points across the 5 major categories of MUG and then

further subdivide category allocations among the different subcategories. This method is consistent with prior consumer behavior research and is termed the constant sum approach (Hawkins et al. 1995). Given that users have different requirements for Web sites from different industries, we expect the usability categories' weights to be determined by the evaluator depending on the product (industry) and task or role (e.g., customer, investor, etc.). The usability criteria weights govern all of the Web sites that an evaluator is assessing for a particular industry.

While the weights are specific to a product (industry) for a particular task or role, they do not represent user evaluations of particular Web sites per se. In the next step, users provide ratings for specific Web sites on various subcategories. The weights and ratings together are then used to assess the overall usability for each site. An example is shown in Table 1. The final number yielded by this computation then constitutes the usability metric.

Table 1 Illustration of the Use of Weights and Ratings in Determining Usability

Categories and Subcategories	Category Weight	Subcategory Weights	Rating (1 to 10)	Weighted Rating	Maximum Rating
Content	45				
Relevance		15	8	120	150
Media use		10	4	40	100
Depth/breadth		10	5	50	100
Current information		10	7	70	100
Ease of use	30				
Goals		15	4	60	150
Structure		10	10	100	100
Feedback		5	5	25	50
Promotion	5	5	10	50	50
Made-for-the-medium	15				
Community		10	8	80	100
Personalization		0	N/A	0	0
Refinement		5	8	40	50
Emotion	5				
Challenge		0	N/A	0	0
Plot		0	N/A	0	0
Character strength		5	7	35	50
Pace		0	N/A	0	0
Overall rating				670	1,000

Notes.

(1) The results shown here illustrate what one user will provide for a particular site. For additional sites in the same industry for the same task, the weights will remain the same but the ratings and, therefore, weighted ratings will differ.

(2) Category weights add up to 100.

(3) The weight assigned to each category is distributed across the various subcategories.

(4) The weighted rating is the product of subcategory weight and the assigned rating.

(5) The maximum rating is the subcategory weight multiplied by 10.

3.1. Instrument Development

The first step in the instrument development process was to ensure content validity (Cook and Campbell 1979, Straub 1989, Venkatraman and Grant 1986). Content validity ensures that the operationalization of a construct adequately represents the domain of coverage of the construct. The typical procedures for assessing content validity are literature reviews, expert assessments, and subjects' assessments. Given the multidimensional nature of MUG, we generated multiple candidate items for each category and subcategory to measure users' weight assessments. Similarly, a broad set and multiple candidate items were generated to measure users' ratings (Diamantopoulos and Winklhofer 2001).

Item refinement and final item selection were accomplished in four phases. In the first phase, two experts in the domain of usability and IS, two experts in measure development and statistical procedures, and two Ph.D. students in IS labeled each item to describe what they believed was measured by the particular item. Most items were labeled consistent with the category and subcategory label. The experts also made suggestions for wording changes. Minor wording changes were made. In the second phase, this procedure was repeated among 40 undergraduate students. There was a high degree of consistency in the labels given by participants for each item. Once again, the labels were consistent with the underlying category and subcategory. After slight additional refinement, a third phase

of labeling was conducted. The convergence and accuracy increased to nearly 100%. Finally, in the fourth phase, one item was chosen to represent each category and subcategory of weight and rating, respectively. A group of 30 randomly chosen individuals with previous Web experience were asked to label the items. Following the high degree of convergence across subjects in this phase, we believed that content validity was established.

An issue that merits attention in this regard is the use of single-item scales for the various category and subcategory weights and ratings. Nunnally (1978) suggests and Venkatraman and Grant (1986) acknowledge that single-item scales are acceptable where the construct being measured is unidimensional. Given that MUG proposes categories such that each category has various subcategories, and the fact that usability dimensions are explicitly modeled via the subcategories, the situation here is well suited to measuring various categories and subcategories via single-item scales.

This choice is particularly important from a pragmatic perspective. For example, consider a case where a company wants a between-subjects assessment of their Web site compared to four competitors. In other words, users are required to provide weight assessments on 5 categories and 14 subcategories (19 weight assessments) for a particular industry and provide ratings on 14 subcategories for 5 Web sites (70 rating assessments). Even with solely single-item measures, this requires each respondent to provide 89 responses in addition to the time it takes to actually surf the Web sites, when necessary, if just to refresh the user's memory. Single-item measures have been used in prior research, specifically when the resulting measure is an index measure that is computed from the measurement of various individual items. This is particularly common in the job descriptive index research (Dwyer and Fox 2000). In the case of index measures, it is more important to ensure that the domain is adequately sampled so none of the possible contributing factors are omitted. Therefore, the most critical validity in the case of such index measures is content validity. Moreover, convergent validity and discriminant validity are not as crucial because index measures typically use formative indicators as opposed to the more common reflective indicators (Bollen 1989, Bollen and Lennox 1991, Diamantopoulos and Winklhofer 2001).

Subsequent to the development of the various items, instructions necessary for the administration of the instrument were written. Experts and peers also evaluated these and wording changes were effected. Table 2 shows the general instructions provided to participants. Also shown in this table are the specific instructions for both the customer and the investor tasks.

To assign weights to various categories, participants were first provided instructions regarding the weighting process. The weights assigned were for a particular *industry* for a particular task. Thus, the weighting scheme would apply across all sites in that industry or product group. The participants were given the 5 categories across which they distributed the 100 points. Table 3 shows this information.

Table 2 General Instructions and Task Instructions

Instructions for Tasks

This survey will ask you to provide the following two sets of information:
(1) The first set of information relates to *how important you believe* several attributes are in determining the usability of any Web site in a particular industry.
(2) The second set of information relates to *how well you believe* several competing sites perform on the various attributes, regardless of how important a particular attribute was.

Customer Task

For all of the activities that you do today, we ask that you play the role of an individual or household customer of the firm. In other words, when you provide us with information regarding the criteria that are important to you, please remember that your assigned role (i.e., the tasks that you perform) is that of an individual or household customer. For example, this means that if you were indicating how important "free stuff" is to you when assessing the usability of Web sites in the insurance industry, remember that you are an individual or household customer.

Investor Task

For all of the activities that you do today, we ask that you play the role of an individual or household investor in the firm. In other words, when you provide us with information regarding the criteria that are important to you, please remember that your assigned role (i.e., the tasks that you perform) is that of an individual or household investor. For example, this means that if you were indicating how important "investor section in a Web site" is to you when assessing the usability of Web sites in the insurance industry, remember that you are an individual or household investor.

Table 3 Instructions Regarding Weighting Scheme and Category Weighting Items

Weights:
You have 100 points to distribute across the 5 categories shown below. You should distribute the points based on the relative importance of the categories in determining the usability of Web sites in _____ industry for your task of the _____ . In other words, the more important a category is to you for sites in the _____ industry for the _____ task, the more points you allocate to it. Note that you are *not* saying how good a particular site is with regard to each category, but rather how *important* each category is to you in deciding the overall usability of Web sites in _____ industry for your task of the _____ . The computer will show you a tally of what is allocated in the column below labeled "Total." You may change the allocation until you finalize it by clicking it on the "Submit" button below.

Category	Explanation	Weight
Content	The extent to which a Web site offers informational and transactional capability.	
Ease of use	The extent to which using a Web site is free of effort.	
Promotion	The extent to which a Web site is well promoted on the Web and other media.	
Made-for-the-medium	The extent to which a Web site can be tailored to fit your specific needs.	
Emotion	The extent to which a Web site evokes emotional reactions from you.	
	Total (Maximum: 100 points)	

Following the assignment of weights, participants were asked to distribute the points assigned to each category across the various subcategories. As discussed earlier, four of the five categories had subcategories. Along with instructions (see Table 4) to distribute the weights, the user was prompted via the items shown in Table 4. Thus, the importance of various usability subcategories would be established for a particular user for a particular task in a particular industry or product group.

Once the participant assigned weights to various categories and subcategories for the assigned task in the industry, they rated various Web sites in terms of their quality on the particular attribute. The subjects browsed through Web sites and provided their ratings for specific sites. A single screen with all categories and subcategories and the 10-point scale (anchored on "extremely poor" and "extremely good") was displayed. When the categories or subcategories were presented, the one-line explanations included earlier (see Table 4) were modified to reflect that the specific rating regarding each Web site was sought. Additional instructions were provided—the instructions pertaining to the customer task are shown in Table 5 (similar instructions were provided for the investor task).

In addition to content validity, it is also critical to assess construct validity. Construct validity is the extent to which the items for a construct present an accurate operationalization of the construct (Cook and Campbell 1979, Straub 1989, Venkatraman and Grant 1986). It is typically assessed via convergent and discriminant validity. Both convergent and discriminant validity are problematic in this case because of the inherent multidimensionality of the proposed usability metric that results in a set of measures that are not expected to converge with each other. Further, because a number of single-item scales are employed to assess weights and ratings, we had to use an alternate method of assessing construct validity that captures the essence of convergent validity. We examine how closely the *calculated usability metric* for a Web site relates to a multi-item scale of overall usability. The higher the correlation between the two constructs, the more accurately the multidimensional calculated usability measure represents the construct of usability. This is in keeping with the spirit behind Campbell's (1960) recommendations for alternative tests of construct validity.

Moreover, this approach is analogous to a multimethod approach for convergent validity, for example, the extent to which two different methods for measuring the same construct yield results that are highly correlated (Campbell and Fiske 1959). To examine overall usability, we used the following three-item scale anchored on "extremely poor" and "extremely good":

- How do you rate the overall usability of the Web site?

Table 4 Instructions and Items for Weight Distribution Across Subcategories

Please allocate the ____ points that you allocated to content, across the following four subcategories of content. You should distribute the points based on the relative importance of the subcategories in determining the usability of Web sites in _____ industry for your task of the _____ . In other words, the more important a subcategory is to you for sites in the _____ industry for the _____ task, the more points you allocate to it. Note that you are *not* saying how good a particular site is with regard to each subcategory, but rather how *important* each subcategory is to you in deciding the overall usability of Web sites in _____ industry for your task of the _____ . The computer will show you a tally of what is allocated in the column labeled "Total" below. You may change the allocation until you finalize it by clicking it on the "Submit" button below.

Subcategory	Explanation	Weight
Category: Content		
Relevance	The extent to which a Web site offers content that is relevant to the core audience.	
Media use	The extent to which a Web site uses media appropriately and effectively to communicate the content.	
Depth and breadth	The extent to which a Web site provides the appropriate breadth and depth of content.	
Current and timely information	The extent to which a Web site provides current and timely information.	
	Total (Maximum: . . . points)	
Instructions same as before were provided.		
Category: Ease of Use		
Goals	The extent to which a Web site offers clear and understandable goals.	
Structure	The extent to which a Web site is well structured and organized.	
Feedback	The extent to which a Web site provides clear and understandable results and feedback regarding your progress.	
	Total (Maximum: . . . points)	
Instructions same as before were provided.		
Category: Made-for-the-Medium		
Community	The extent to which a Web site offers you the opportunity to be part of an online group or community.	
Personalization	The extent to which a Web site can treat you as a unique person and respond to your specific needs.	
Refinement	The extent to which a Web site reflects the most current trend(s) and provides the most current information.	
	Total (Maximum: . . . points)	
Instructions same as before were provided.		
Category: Emotion		
Challenge	The extent to which a Web site offers you an element of challenge.	
Plot	The extent to which a Web site provides an interesting story line.	
Character strength	The extent to which a Web site ties to individuals, within and outside the organization, who have credibility.	
Pace	The extent to which a Web site allows you to control the pace at which information you interact with it.	
	Total (Maximum: . . . points)	

Table 5 Instructions for Rating Web Sites

When you rate each Web site for various attributes, please rate them from the perspective of an individual or household customer of the firm. For example, if you are rating a Web site on "free stuff," provide your rating on how much or how good the "free stuff" is on the Web site from your perspective as an individual or household customer. So "free stuff" given to commercial clients will not count. Also, when you provide ratings, the weights you assigned in the first several steps do not play a role. So it does not matter whether you thought "free stuff" was important to you or not, you are simply providing a rating of the "free stuff" made available to the individual customer.

- How do you rate the overall design of the Web site?
- How do you rate your overall experience at the Web site?

Two pilot studies were conducted to evaluate the viability of the research approach and proposed procedures. The objective of the first pilot study was to test the overall applicability of MUG and also to more closely examine the method of assigning weights and ratings to arrive at usability evaluations. Instructions

to be used in the actual study were also refined in this pilot study.

The first study involved 80 seniors enrolled in an electronic commerce design and development class. As part of the course, students read a white paper from Microsoft's Web site that describes MUG. The instructor of the class, one of the authors, described MUG and the weighting and rating methodology with examples in two lecture meetings, lasting 2 hours and 30 minutes. The students were expected to learn the concepts from the perspective of quizzes, exams, and assignments and projects. Specifically, two quizzes and the midterm exam featured questions related to MUG. The assignments and projects required that the students evaluate three sets of Web sites over the course of a semester: (1) any five Web sites from a single industry chosen by the student, (2) three auto manufacturers and three auto sales intermediaries, and (3) the Web site of the school by choosing one of many roles—current student, prospective undergraduate student, prospective graduate student, graduating student seeking employment, etc. In all three cases, students provided detailed justification for their decisions, thus providing the researchers with information that contributed to the development and refinement of the instrument.

A second pilot study was conducted to examine the pragmatic aspects of the field study to be conducted, including the working of the kiosk and the server used to support the kiosk; in addition, the second pilot study was to serve as a field test of the instrument. The second pilot study included the actual survey instrument and Web sites being studied and was conducted during a three-hour period on a Saturday at one of the participating locations. A total of 104 participants were involved in this pilot study. All practical aspects related to the technology worked well. Further, an analysis of the data revealed acceptable levels of reliability and validity, prompting the researchers to continue with the large-scale field study as planned.

3.2. Participants

The population of interest in this study was Internet users. The sampling frame was visitors to three branches of a major retail store during a three-day period (Friday, Saturday, and Sunday). The specific participants were identified using a "mall intercept," which is discussed in greater detail in the "Procedure" subsection. A total of 1,823 individuals agreed to participate in the study and 1,475 provided usable responses, for an effective response rate of 81%; 527 of the 1,475 participants were women (35.7%). Unusable responses are primarily attributable to incomplete information. Based on information from store employees who supervised the data collection, some participants found the survey to be too lengthy and/or were so completely unfamiliar with the Web sites that they were visiting that they felt unable to form assessments regarding various aspects of the Web sites in a short time frame—in all of these cases, participants did not complete the survey. In our view, it was better, in fact, for such participants to withdraw from the study rather than to provide inaccurate responses.

3.3. Procedure

The participants were recruited at three branch locations of a major electronics retail store. In the mall-intercept method, individuals are invited to participate in the study. Specifically, a promotion desk staffed by a store employee was set up to facilitate active participant recruitment with a "request to complete a survey and get a $10 gift card." The $10 incentive could be used for purchasing store merchandise and had no expiration date. Three kiosks were set up in the participating stores for the participants to browse the specific Web sites that were being studied and also respond to the questionnaire. The use of three kiosks at each store helped to minimize participant wait time. In cases where there was a wait time, participants were given a radio device that could be paged when a kiosk became available, thus allowing participants to shop instead of having to wait—such an approach was seen as a way of enhancing response rate.

Web sites were chosen from four industries: airlines, bookstores, auto manufacturers, and car rental agencies. The specific Web sites chosen from the different industries are shown in Table 6. Demographic information about the participants broken down by the various products (airline, bookstore, auto manufacturer, and car rental) and tasks (customer and investor) is given in Table 7. Within each product and task are the sample size, number of women/men, age, and income.

When a participant arrived at a kiosk, he or she was

Table 6 Web Sites Studied

Airline	Bookstore	Auto Manufacturer	Car Rental
American Airlines	Amazon.com	BMW	Alamo
Delta Airlines	BarnesandNoble.com	Chrysler	Avis
Northwest Airlines	Booksense.com	Ford	Budget
United Airlines	Borders.com	GM	Hertz
US Airways	VarsityBooks.com	Mercedes	National
		Porsche	

Note. To protect the anonymity of these sites, and organizations that they represent, sites are listed here alphabetically and do not represent the order in which they were entered in the data file.

prompted with a request to fill out a survey regarding Web sites in one of the four industries. The specific industry assigned to a particular participant was chosen randomly by the computer. The participant was also randomly assigned a task—customer or investor—by the computer, and presented the instructions for the specific task (see Table 2). The respondent then provided his or her perception of the relative importance (weights) of the different criteria. Next, additional background information from the participants was collected.

Following this, the participants visited the Web sites. The order of presentation of the different Web sites was randomized by the computer. Every participant was given five minutes to browse each Web site—the system prompted them if they wished to continue to browse the Web site after that point and did allow for further browsing. Further browsing was deemed acceptable and important because it was possible for participants to require more information about a Web site (and the organization) as they could have been minimally familiar with one or more sites. After browsing each Web site, the participants responded to a three-item questionnaire regarding the overall usability of the Web site. Next, participants rated different MUG attributes for the site on a 10-point scale. Finally, demographic information was gathered. Based on system logs, it was determined that the average time spent browsing the Web sites was about 22 minutes and the average time spent filling out the survey was about 25 minutes.

4. Results

The multi-item usability scale was found to be highly reliable with Cronbach αs being over 0.80 in all 21 cases; i.e., for each of the 21 sites studied in this research. Because the weights and ratings were measured via single-item scales, much of the validity was already established through the careful procedures undertaken earlier. However, one important step was to examine the correlation between the calculated usability metric and the three-item measure. Table 10 reports the correlations along with the descriptive statistics for all sites studied in the current research. The correlations between the calculated usability rating and usability measured using the three-item scale were very high—ranging from 0.71 to 0.93 across the various products and tasks, providing strong evidence of convergent validity (Campbell and Fiske 1959).

The realism of assigned tasks—customer and investor—was tested via two items on a seven-point scale. The items were: (1) "The task I have been asked to do is consistent with what I might do at the Web sites of these companies" and (2) "The task I have been asked to do is a realistic representation of what I might do at the Web sites of these companies." The results revealed that participants did, indeed, consider those tasks to be realistic and something that they would typically do at the Web site of the companies (Customer: $M = 6.1$, $SD = 0.52$; Investor: $M = 5.9$, $SD = 0.64$).

4.1. User Assessment of Weights

Means and standard deviations of the weights of the different categories broken down by product and task are shown in Table 8. The sample for this data analysis was 1,475 because each participant's response provided one set of weights.

ANOVAs followed by Scheffe's tests (Neter et al. 1985) were conducted to inspect differences in weights across industries and tasks. The following interesting findings emerged:

- Content was the most important category in all eight groups (four products, two tasks). Customers of all products deemed the content of a Web site to be equivalently important. Investors believed content to be more important than did customers.
- The second category of ease of use was modestly

Table 7 Demographic Characteristics Broken Down by Product (Industry) and Task (Customer vs. Investor)

	Customer				Investor			
Industries	N	M/F	Age	Income	N	M/F	Age	Income
Airline	230	151/79	30.07 (6.62)	57,004 (8,275)	211	130/81	31.15 (6.91)	59,645 (8,275)
Bookstore	177	111/66	30.01 (6.64)	58,145 (8,045)	148	97/51	29.87 (6.88)	58,822 (8,095)
Auto manufacturer	201	130/71	31.07 (6.84)	57,987 (7,922)	227	155/72	30.17 (6.19)	58,727 (8,001)
Car rental	149	95/54	30.98 (6.80)	56,545 (7,887)	132	79/53	29.93 (6.17)	58,888 (7,572)

Note. Mean and standard deviation pairs are indicated for age and income. Age is in years and income in dollars per annum. The numbers in parentheses are standard deviations.

important across all eight groups. Customers deemed ease of use to be more important than investors.

• The importance of promotion varied across tasks, regardless of industry—while somewhat important to customers, it was weighted nearly twice as much by investors. Thus, the determinants of promotion followed a pattern similar to content.

• Made-for-the-medium was influenced by a two-way interaction of product and task. It appears that made-for-the-medium was more important to customers in three of the four industries, with auto manufacturing being the exception. The role of made-for-the-medium was relatively stable across all products in the investor task.

• Emotion was also influenced by a two-way interaction of product and task. In contrast to made-for-the-medium, in this case, the auto manufacturing sites' customers deemed emotion to be very important compared to customers in all industries. However, investors viewed emotion to be only minimally important.

4.2. User Assessment of Usability

After understanding user decisions regarding the weights, we then examined overall usability. Recall that each participant provided a usability rating for each site in the industry, thus resulting in multiple responses per participant. Table 9 shows the sample in each of the different groups broken down by product and task.

The means and standard deviations of the usability ratings (calculated and rated) of the different sites broken down by industry and task and the correlations between these usability ratings are shown in Table 10. As before, ANOVAs followed by Scheffe's tests were conducted to examine differences within sites across tasks, sites for a specific role, and industries both within a role and across roles.

A comparison across industries reveals that bookstore sites scored highest with both customers and investors. One of the noteworthy observations is that, by and large, most sites were seen to be higher in terms of usability by customers when compared to investors. In addition, the following specific findings emerged:

• Within the airline industry, usability for the customer task showed the least variability. For the investor task, participants rated all sites to be equivalent in terms of usability, but these ratings (low fives) were significantly lower than even the lowest rating for the customer task.

• The bookstore sites were the best and the worst across all sites in all four industries for the customer task, showing greatest variability. Interestingly, the bookstore sites that were rated lower for the customer task emerged as being quite highly rated for the investor task. The best sites for the customer task were also found to be fairly good in terms of usability for the investor task, but the usability ratings for the investor task were much lower than the ratings for the customer task.

• The auto manufacturer sites' usability for the customer task exhibited a great deal of variance, although not as much as the bookstore sites. The investor ratings of usability were lower than customer ratings for five out of the six sites, with usability scores in the fives and sixes. Also, five out of six sites were rated very

Table 8 Relative Importance (Weights) of Different Attributes Across Products (Industries) and Tasks (Customer vs. Investor)

		Airline		Bookstore		Auto Manufacturer		Car Rental	
Assigned Tasks	Categories	M	SD	M	SD	M	SD	M	SD
Customer	**Content**	32.8	5.67	33.2	6.02	38.1	8.10	33.2	6.31
	Ease of use	16.4	7.08	15.0	6.53	12.7	5.72	15.8	6.22
	Promotion	10.1	4.21	12.2	4.04	10.9	4.30	13.1	4.51
	Made-for-the-medium	32.4	8.08	30.4	7.63	14.5	6.10	29.1	7.02
	Emotion	8.3	3.77	9.2	3.61	23.8	10.87	8.8	3.21
	Total	100		100		100		100	
Investor	**Content**	40.2	7.72	41.4	7.13	46.6	7.14	39.8	6.68
	Ease of use	11.4	4.10	13.2	4.13	9.5	3.32	15.0	4.01
	Promotion	20.3	5.71	18.7	4.89	19.5	5.16	20.1	5.19
	Made-for-the-medium	25.3	8.82	22.9	7.22	20.8	6.67	24.4	6.99
	Emotion	2.8	0.45	3.8	0.33	3.6	0.81	0.7	0.21
	Total	100		100		100		100	

Table 9 Sample Size for Usability Data Analysis

Industries	Customer	Investor	Total
Airline	1,150	1,055	2,205
Bookstore	885	740	1,625
Auto manufacturer	1,206	1,362	2,622
Car rental	745	660	1,405
Overall	3,986	3,817	7,803

close to each other on the investor task, with one site being much worse.

- The results for the Car rental sites were similar to the airline industry for the customer task—there was about a one-point difference in the usability scores across sites, with scores ranging from about six to seven. Similar to the auto manufacturer sites, the ratings of investors were lower than those of customers; the ratings by investors varied greatly across sites. The car rental sites showed the poorest usability across all sites studied for the investor task.

5. Limitations

Prior to discussing the implications of our work, certain limitations of the research that influence the interpretation of findings must be acknowledged. The sampling method used here could have inadvertently introduced some selection bias in the choice of participants, although the large sample size and the fact that data were collected over three locations and three days does introduce a greater degree of randomness in sample selection. Subject motivation is potentially an issue here as $10 is a modest incentive. However, the fact that participants chose to browse sites for an average of 22 minutes suggests that they were sufficiently engaged in the task to provide meaningful responses.

Because of resource constraints, we elected to study a small number of industries and, therefore, the generalizability of the findings to other industries needs to be investigated in future research. It is possible that in spite of the fact that between five and six sites were selected for each industry (with the goal of obtaining broad representation across the major firms in each industry), there may be an overall industry effect related to the quality of the Web site. In other words, one industry or another may have significantly poorer interfaces. Although we did not explicitly examine this in our study, it is a useful avenue for future research. A systematic bias in this result might impact the interpretation of generalizable assessments across industry and across task. Additionally, we investigated usability in the context of business-to-consumer sites. The extent to which these findings generalize to business-

Table 10 Usability Ratings by Product and Task

Industries	Assigned Tasks	Usability Ratings	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6
Airline	**Customer**	**Calculated**	7.22 (2.12)	6.82 (2.09)	7.10 (2.08)	6.32 (2.22)	6.17 (2.10)	
		Rated	7.10 (2.07)	6.53 (2.01)	6.99 (2.08)	6.10 (1.97)	5.99 (1.87)	
		Correlation	0.87***	0.82***	0.84***	0.90***	0.81***	
	Investor	**Calculated**	5.17 (1.33)	5.08 (1.08)	5.10 (1.17)	5.07 (1.02)	5.02 (1.21)	
		Rated	5.10 (1.11)	5.01 (1.21)	4.92 (1.03)	4.90 (1.07)	5.02 (1.08)	
		Correlation	0.88***	0.92***	0.78***	0.81***	0.93***	
Bookstore	**Customer**	**Calculated**	8.12 (1.05)	8.07 (1.02)	5.20 (2.02)	4.10 (2.08)	4.08 (1.87)	
		Rated	8.19 (1.03)	8.11 (1.02)	5.89 (1.88)	4.04 (2.11)	4.14 (1.82)	
		Correlation	0.91***	0.92***	0.71***	0.90***	0.91***	
	Investor	**Calculated**	6.21 (1.14)	7.11 (1.09)	6.97 (1.09)	7.91 (1.45)	7.12 (1.22)	
		Rated	6.23 (1.13)	7.06 (1.31)	7.01 (1.33)	7.96 (1.39)	7.10 (1.09)	
		Correlation	0.82***	0.80***	0.83***	0.87***	0.87***	
Auto manufacturer	**Customer**	**Calculated**	7.93 (1.41)	7.03 (1.12)	7.71 (1.21)	6.83 (1.12)	7.82 (1.23)	5.89 (2.22)
		Rated	7.84 (1.07)	7.10 (1.09)	7.67 (1.19)	6.90 (1.07)	7.72 (1.09)	5.79 (2.19)
		Correlation	0.86***	0.88***	0.91***	0.86***	0.84***	0.91***
	Investor	**Calculated**	6.62 (1.08)	5.02 (1.07)	6.66 (1.14)	6.79 (1.03)	6.82 (1.03)	6.83 (1.01)
		Rated	6.81 (1.11)	5.21 (1.17)	6.67 (1.19)	6.89 (1.11)	6.71 (1.22)	6.72 (1.21)
		Correlation	0.80***	0.82***	0.88***	0.84***	0.84***	0.87***
Car rental	**Customer**	**Calculated**	7.22 (1.08)	7.11 (0.98)	6.10 (1.34)	6.44 (1.20)	6.22 (1.28)	
		Rated	7.71 (1.24)	7.10 (0.91)	6.12 (1.30)	6.38 (1.22)	6.27 (1.22)	
		Correlation	0.79***	0.94***	0.91***	0.88***	0.90***	
	Investor	**Calculated**	5.11 (1.07)	4.13 (1.11)	3.41 (1.08)	3.44 (1.01)	3.22 (0.98)	
		Rated	5.08 (1.04)	4.10 (1.17)	3.32 (1.30)	3.41 (0.91)	3.36 (0.87)	
		Correlation	0.83***	0.78***	0.86***	0.87***	0.89***	

Note. Usability scores are on a 10-point scale. The number in parentheses is the standard deviation.

to-business electronic commerce sites would require further research.

Participants were assigned the role of a customer or investor—while most participants could easily fit into the customer role, it is possible that the kiosk in the store did not create a natural online shopping environment. Also, it is possible that not all participants assigned the investor role may have felt comfortable in the role. However, this limitation is somewhat alleviated in today's world of significant direct investment by individuals via electronic brokers such as E*Trade and Ameritrade (see Modahl 2000). In any case, future research should address this limitation with a more realistic sample for the investor task; more broadly, research should investigate other tasks to examine the generalizability of the current findings and further establish the validity of this instrument and the associated metric. Given the scale and manner of data collection, logistical constraints dictated that the survey instrument be of reasonable length and, therefore, we were not able to utilize multi-item measures for all the constructs. Finally, we did not measure actual behavior, and while there is sufficient prior evidence to suggest that perceptions of usability result in certain desirable behaviors, our research did not specifically test this relationship.

6. Discussion and Implications

Our goal in this paper was to describe a metric and procedure for assessing the quality of an organization's Web presence. Arguing that the usability of a Web site is a fundamental component of the total user experience, and motivated by evidence suggesting that

consumer electronic commerce is likely to be of significant strategic importance to companies in the digital economy, we suggested that such a metric would be useful for both researchers and practitioners. Field application of the metric revealed that not only does the instrument demonstrate good psychometric properties, but the evaluation procedure also provides detailed insight into the relative importance of specific aspects of Web design for different types of users across different types of industries. As such, this information is likely to be invaluable in helping improve the design of a Web site. Below we discuss some of the more interesting findings and their implications for both research and practice.

6.1. The Drivers of Web Site Usability

As argued earlier, a critical requirement of a useful usability metric is that it demonstrates the ability to discriminate across Web sites from different industries and among different types of users. Our results revealed that the metric is indeed able to do so: weights assigned by participants to different MUG categories and subcategories suggest that the salience of usability characteristics varied depending on the user task and industry to which the Web site belonged. Echoing findings from a large body of technology acceptance research (e.g., Moore and Benbasat 1991, Venkatesh and Davis 2000), we thus found that the instrumental goals of a user are key determinants to what they seek from a Web site. The importance of content was highest across all attribute categories, consistent with the observation made by Cole et al. (2000), suggesting that the relevance of substantive information contained on a Web site, its completeness as assessed by information depth and breadth, and its currency are all critical to the Web site visitor. We also found that the importance of content was contingent upon the task that the visitor was trying to accomplish: consumers rated content as significantly more salient than investors, presumably because they intended to use the Web site to fill a specific consumption need.

Findings reveal that the ability of a Web site to support promotion is more important for investors than it is for consumers. This result possibly relates to investors funneling their resources into investment opportunities that are more likely to generate greater returns. Therefore, investors' appraisal of advertising and promotion quality will contribute to their assessment of a firm's future success. The possibilities offered by the Web environment as an advertising and communication medium are unprecedented (Hoffman and Novak 1996), and a corporate Web site can be a powerful means to building a strong brand image. Indeed, for the pure play Internet company that does not have a "brick and mortar" counterpart, the ability of the Web site to "promote" the company is likely to be a crucial factor in the company's success. In the pre-Internet era, information flow to investors was closely controlled through corporate communications groups and standard public disclosures such as annual reports. With online investing (as well as for investments in online companies) investors will increasingly evaluate firm potential through information disseminated on the company Web site, and the results of the current research show that a key component of their evaluation will be the extent to which the Web site supports promotion.

Several scholars have alluded to the ability of hypermedia environments in general, and the Web in particular, to engender emotional responses among users (Hoffman and Novak 1996, Agarwal and Karahanna 2000). Variously labeled as the "flow" experience, "cognitive engagement," and "cognitive absorption," the central notion here is that the multimedia capabilities, richness, and interactivity of the Web environment have the potential to engage users in ways not exhibited by other media. In recognition of this capability, one of the key MUG categories pertains to the extent to which a Web site generates emotion when users interact with it through judicious use of features such as character strength and pace.[1] We found that the importance of emotion was contingent on both task and product characteristics. In particular, consumers were more concerned than investors with regards to the ability of a Web site to appeal to their emotions. This attribute was especially important to the auto manufacturers' Web sites, where consumers assigned

[1]As presented in the discussion of MUG categories and subcategories, character strength relates to the credibility conveyed by the site, particularly via the individuals portrayed on the site, while pace is the extent to which the site provides users an opportunity to control the flow of information.

it 24 points on average when compared to less than 10 for any of the 3 other product types examined. Autos represent high involvement, high outlay, nonrepeat purchases, and are a purchase decision that a consumer will have to live with for a reasonably extensive time period, thus explaining why the consumer is emotionally engaged with the product.

Additional unique characteristics of the Web environment, such as its ability to support community, personalization, and continual refinement; i.e., the "made-for-the-medium" category from MUG, were similarly affected by a product-task interaction. In contrast to emotional appeal, here consumers rated this attribute (made-for-the-medium) as being least important for auto manufacturers, while investors did not appear to believe that this attribute discriminated in usability across product types. One plausible explanation for this finding rests in the essential difference in the functional activity supported by the Web sites examined here. Hoffman et al. (1995) used the dimension of commercial activity supported to develop a functional typology of Web sites consisting of online storefronts, Internet presence sites, content sites, malls, incentive sites, and search agents. In their typology, the airline, car rental, and bookstore sites all represent online storefronts, while the auto manufacturer sites are fundamentally Internet presence sites that provide information but do not have transaction capabilities. Moreover, auto purchases are typically made through intermediaries such as dealers and rarely directly through the auto manufacturers. It appears that the consumer distance from the seller in the case of autos in the physical world is being transferred to cyberspace, and while consumers want to use the auto manufacturer Web sites to evaluate competing products, they do not appear to view the seller as one with whom they would want to establish a deep relationship through community and personalization. On the other hand, airlines, book sellers, and car rental agencies are more likely to be used for repeat purchases, thereby rendering the made-for-the-medium attributes more salient.

6.2. Implications for Practice

In this research, we propose and operationalize a new heuristic evaluation method for assessing Web site usability. The result is also a new set of metrics. HCI researchers and practitioners have long been concerned about the relative effectiveness and efficiency of alternative usability evaluation methods. Our findings suggest that this method provides detailed insight into potential design defects, especially as it focuses on the relative importance of the various categories and subcategories of usability. However, the high correlation between the computed usability measure (the weighted sum of products and ratings) and the direct overall usability measure points to some interesting choices that HCI practitioners can make. If the goal of usability evaluation is simply to measure how usable a specific Web site is, then the simpler overall usability scale provides an efficient method of operationalizing such an assessment. On the other hand, if the goal is to isolate specific design defects, the detailed procedure employed here is likely to be valuable. In general, both approaches are recommended as they complement each other through different stages of the continual usability evaluation life cycle.

Two key implications for and contributions to practice emerge from the findings discussed above. First, to the extent that usability is an important metric for assessing Web site design, managers need systematic methodologies for performing usability assessments. Although the face validity of the MUG criteria is indisputable, the guidelines do not include a method for operationalizing assessments. We presented such a method, developed an instrument to capture the various criteria, and demonstrated the feasibility of the approach through an extensive field study. A second overarching implication is the fact that users who visit a Web site do so with a variety of goals, predispositions, and purposes in mind. With the increasing importance of marketing initiatives such as personalization and dynamic content, our results shed light on what factors need attention: product-task interactions do exist and must be focused on in Web site design.

6.3. Implications for Research

For researchers, we have provided initial insight into factors that are likely to be significant antecedents of Web site usability. The procedure we describe is extensible to other industries and types of users. However, several areas for fruitful future research remain. Our focus was on instrument development and validation, and although we showed that the instrument produces

differential weights and ratings depending on the user's task and the industry to which a Web site belongs, we did not develop theoretical arguments in support of these effects. From the perspective of theory development, there are opportunities to offer rich explanations of user assessments of Web sites by synthesizing research from multiple streams, including marketing, information technology acceptance, and human-computer interaction. For instance, marketers have often used multiple user characteristics and demographic variables such as age, income, and gender as explanatory variables for purchase decisions (e.g., Hawkins et al. 1995). Future research may consider incorporating these characteristics along with task and product to examine how they influence user assessments of usability. While the link between usability and actual behavior is implicit in all of our arguments, we did not specifically test it here. Empirical testing of this link would be a logical next step in extending this research.

The research presented here also needs to be extended across more products and industries to determine the robustness of the instrument. Although we used an alternative measure of usability (a three-item rating) in addition to the multidimensional measure, it would be useful to also compare the multidimensional measure with ratings provided by third parties or some other measurement technique. In this way, we can be assured that there are no systematic differences in industries that are not attributable to the sampling procedure here. For HCI researchers, a fruitful area for future work would be to compare weights and ratings assigned by actual users to those provided by usability experts. Finally, controlled experiments would support the field study approach followed here by reducing extraneous variance associated with other factors, thereby increasing confidence in the findings.

7. Conclusion

The Internet in general and e-commerce, in particular, exhibit the characteristics of disruptive technologies and processes in that they offer the promise and threat of fundamentally altering how business is conducted. Managers and firms desirous of exploiting the opportunities offered by these changes need to continually assess if their investments are yielding desired returns. In this paper, we examined one such investment, the design of a corporate Web site, and offered a metric; i.e., usability and a procedure for operationalizing its use. The metric exhibits good psychometric properties and provides detailed insight into specific design elements that need attention. As net-enabled organizations continue to increase investment in their Web presence, the results presented here should be useful in an on-going assessment of potential impact. In sum, the current research contributes an important metric to help managers understand and predict the likely success of e-commerce.

Acknowledgments

The authors gratefully thank the Senior Editor and three anonymous referees whose comments helped improve the paper.

References

Agarwal, R., E. Karahanna. 2000. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quart.* **24** 665–694.

——, J. Prasad. 1997. The role of innovation characteristics and perceived voluntariness in the acceptance of information technologies. *Decision Sci.* **28**(3) 557–582.

——, P. De, A. Sinha. 1999. Comprehending object and process models: An empirical study. *IEEE Trans. Software Engrg.* **25** 541–556.

——, V. Sambamurthy, R. Stair. 2000. The evolving relationship between general and specific computer self-efficacy: An empirical assessment. *Inform. Systems Res.* **11** 418–430.

Blatt, L. A, J. F. Knutson. 1994. Interface design guidance systems. J. Nielsen and R. L. Mack, eds. *Usability Inspection Methods*. John Wiley and Sons, New York, 351–384.

Bollen, K. 1989. *Structural Equations with Latent Variables*. John Wiley and Sons, New York.

——, K. R. Lennox. 1991. Conventional wisdom on measurement: A structural equation perspective. *Psych. Bull.* **110** 305–314.

Breitenbach, C. S., D. C. Van Doren. 1998. Value-added marketing in the digital domain: Enhancing the utility of the Internet. *J. Consumer Marketing* **15** 558–575.

Campbell, D. T. 1960. Recommendations for APA test standards regarding construct, trait, discriminant validity. *Amer. Psychologist* **15** 546–553.

——, D. W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psych. Bull.* **56** 81–105.

Card, S. K., T. P. Moran, A. Newell. 1983. *The Psychology of Human-Computer Interaction*. Erlbaum, Hillsdale, NJ.

Chau, P. Y. K., G. Au, K. Y. Tam. 2000. Impact of information presentation modes on online shopping: An empirical evaluation of a broadband interactive shopping service. *J. Organ. Comput. Electronic Commerce* **10** 1–22.

Cole, M., R. M. O'Keefe, H. Siala. 2000. From the user interface to the consumer interface. *Inform. Systems Frontiers* **1** 349–361.

Cook T. D., D. T. Campbell. 1979. *Quasi Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin, Boston, MA.

Davis, F. D., R. P. Bagozzi, P. R. Warshaw. 1989. User acceptance of computer technology. *Management Sci.* **35**(8) 982–1003.

Day, G. S. 2000. Managing market relationships. *J. Acad. Marketing Sci.* **28** 45–54.

Diamantopoulos, A., H. M. Winklhofer. 2001. Index construction with formative indicators: An alternative to scale development. *J. Marketing Res.* **38** 269–277.

Dwyer, D. J., M. L. Fox. 2000. The moderating role of hostility in the relationship between enriched jobs and health. *Acad. Management J.* **43** 1086–1096.

Eighmey, J., L. McCord. 1998. Adding value in the information age: Uses and gratifications of sites on the World-Wide Web. *J. Bus. Res.* **41** 187–194.

Gehrke, D., E. Turban. 1999. Determinants of successful Web site design: Relative importance and recommendations for effectiveness. *Proc. 31st Hawaii Internat. Conf. Inform. Systems*, Maui, HI.

Gray, W. D., M. C. Salzman. 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction* **13** 203–261.

Hawkins, D. I., R. J. Best, K. A. Coney. 1995. *Consumer Behavior: Implications for Marketing Strategy*. Richard D. Irwin, Chicago, IL.

Hoffman, D. L., T. P. Novak. 1996. Marketing in hypermedia computer-mediated environments: Conceptual foundations. *J. Marketing* **60** 50–68.

——, ——. 2000. How to acquire customers on the Web. *Harvard Bus. Rev.* **78** 179–184.

——, ——, P. Chatterjee. 1995. Commercial scenarios for the Web: Opportunities and challenges. *J. Comput. Mediated Comm.* **1**.

Jeffries, R. J., J. R. Miller, C. Wharton, K. M. Uyeda. 1991. User interface evaluation in the real world: A comparison of four techniques. *Proc. CHI on Human Factors in Comput. Systems*. New Orleans, LA, ACM, New York, 119–124.

Kantner, L., S. Rosenbaum. 1997. Usability studies of WWW sites: Heuristic evaluation vs. laboratory testing. *Proc. SIGDOC*, Snowbird, UT, 153–160.

Karat, J. 1997. Evolving the scope of user-centered design. *Comm. ACM* **40** 33–38.

Katerattanakul, P., K. Siau. 1999. Measuring information quality of Web sites: Development of an instrument. P. De and J. DeGross, eds. *Proc. 20th Internat. Conf. Inform. Systems* Charlotte, NC, 279–285.

Keeker, K. 1997. Improving Web-site usability and appeal: Guidelines compiled by MSN usability research. ⟨http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnsiteplan/html/improvingsiteusa.asp⟩.

Keevil, B. 1998. Measuring the usability index of your Web site. *Sixteenth Annual ACM Internat. Conf. Comput. Documentation*, Quebec, Canada, September 24–26, 271–277.

Klein, L. R. 1998. Evaluating the potential of interactive media through a new lens: Search versus experience goods. *J. Bus. Res.* **41** 195–203.

Lam, J. C. Y., M. K. O. Lee. 1999. A model of Internet consumer satisfaction: Focusing on the Web-site design. *Proc. 5th Amer. Conf. Inform. Systems*, Milwaukee, WI, 526–528.

Lecerof, A., F. Paterno. 1998. Automatic support for usability evaluation. *IEEE Trans. Software Engrg.* **24** 863–887.

Lohse, G., P. Spiller. 1999. Internet retail store design: How the user interface influences traffic and sales. *J. Comput. Mediated Comm.* **5**.

Mack, R. L., J. Nielsen. 1994. Executive summary. J. Nielsen and R. L. Mack, eds. *Usability Inspection Methods*. John Wiley and Sons, New York, 1–24.

Modahl, M. 2000. *Now or Never: How Companies Must Change Today to Win the Battle for Internet Customers*. Harper Collins, New York.

Moore, G., I. Benbasat. 1991. Development of an instrument to measure the perceptions of adopting an information technology innovation. *Inform. Systems Res.* **2** 192–222.

Neter, J., W. Wasserman, M. H. Kutner. 1985. *Applied Linear Statistical Models*. Richard D. Irwin, Homewood, IL.

Nielsen, J. 1994. Heuristic evaluation. J. Nielsen and R. L. Mack, eds. *Usability Inspection Methods*. John Wiley and Sons, New York, 25–62.

——. 2000. *Designing Web Usability*. New Riders, Indianapolis, IN.

Nunnally, J. C. 1978. *Psychometric Theory*. McGraw Hill, New York.

Olsson, C. 2000. The usability concept reconsidered: A need for new ways of measuring real Web use. *Proc. IRIS 23*, Laboratorium for Interaction Technology, University of Trollhättan Uddevalla.

Payne, S. J., H. Squibb, A. Howes. 1990. The nature of device models: The yoked state hypothesis and some experiments with text editors. *Human-Comput. Interaction* **5** 415–444.

Peppers, D., M. Rogers. 1999. *Enterprise One to One: Tools for Competing in the Interactive Age*. Doubleday, New York.

Polson, P. G., C. Lewis, J. Ripman, C. Wharton. 1992. Cognitive walkthroughs: A method for theory-based evaluation of use interfaces. *Internat. J. Man-Machine, Stud.* **36** 741–773.

Shneiderman, B. 1980. *Software Psychology: Human Factors in Computer and Information Systems*. Winthrop, Cambridge, MA.

——. 1998. *Designing the User Interface*, 3rd ed. Addison-Wesley Longman, Inc., Boston, MA.

Sime, M. E., T. G. Green, D. J. Guest. 1973. Psychological evaluation of two conditional constructions used in programming languages. *Internat. J. Man-Machine Stud.* **5** 105–113.

Straub, D. 1989. Validating Instruments in MIS Research. *MIS Quart.* **13** 147–169.

——, R. Watson. 2001. Transformational issues in researching IS and net-enabled organizations. *Inform. Systems Res.* **12** 337–345.

Subramaniam, C., M. J. Shaw, D. M. Gardner. 2000. Product marketing and channel management in electronic commerce. *Inform. Systems Frontiers* **1** 363–379.

Venkatesh, V. 2000. Determinants of perceived ease of use: Integrating perceived behavioral control, computer anxiety and enjoyment into the technology acceptance model. *Inform. Systems Res.* **11** 342–365.

——, F. D. Davis. 1996. A model of the antecedents of perceived ease of use: Development and test. *Decision Sci.* **27** 451–481.

——, ——. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Sci.* **46** 186–204.

——, C. Speier. 2000. Creating an effective training environment for enhancing telework. *Internat. J. Human-Comput. Stud.* **52** 991–1005.

Venkatraman, N., J. H. Grant. 1986. Construct measurement in organizational strategy research: A critique and proposal. *Acad. Management Rev.* **11** 71–87.

Webster, J., H. Ho. 1997. Audience engagement in multimedia presentations. *Data Base for the Adv. Inform. Systems* **28** 63–77.

Wharton, C., J. Rieman, C. Lewis, P. Polson. 1994. The cognitive walkthrough method: A practitioner's guide. J. Nielsen and R. L. Mack, eds. *Usability Inspection Methods.* John Wiley and Sons, New York, 105–140.

Wixon, D., S. Jones, L. Tse, G. Casady. 1994. Inspections and design reviews: Framework, history, and reflection. J. Nielsen and R. L. Mack, eds. *Usability Inspection Methods.* John Wiley and Sons, New York, 77–104.

Detmar Straub, Senior Editor. This paper was received on January 10, 2000, and was with the authors 11 months for 3 revisions.