AUTOMATICALLY IDENTIFYING USABILITY PROBLEMS IN E-COMMERCE WEBSITES

Jungpil Hahn
Assistant Professor of Management
Krannert School of Management, Purdue University
West Lafayette, IN 47907
Email: jphahn@mgmt.purdue.edu

Robert J. Kauffman
Director, MIS Research Center, and
Professor and Chair, Information and Decision Sciences
Carlson School of Management, University of Minnesota
Minneapolis, MN 55455
Email: rkauffman@csom.umn.edu

Last revised: October 20, 2005

ABSTRACT

Managers at e-commerce firms are in need of proven methods for evaluating the usability of their websites. So, one of the most pressing issues is whether the design of their online storefronts is effective, and if not, which areas require attention and improvements. However, current usability evaluation methods (e.g., user testing, inspection, inquiry) are not well suited to the task at hand. This paper proposes a new approach to website evaluation, which is grounded in the *economic theory of production*. We conceptualize human-computer interaction during online shopping as an economic production process in which customers are making use of various functionalities provided by the e-commerce website in order to complete a purchase transaction. This view enables us to formulate a novel perspective on website usability — the ability to transform inputs (i.e., use of website functionalities) into outputs (i.e., completed purchase transactions). We use *data envelopment analysis* (DEA) as the methodological vehicle for evaluating e-commerce websites and propose two new metrics, *Inefficiency Breadth* and *Unit Inefficiency* to help automatically identify website functionalities that are potentially problematic. The value of the proposed method is illustrated by applying it to the evaluation of a real-world e-commerce website.

Keywords and phrases: Usability evaluation automation, website usability, B2C e-commerce, data envelopment analysis, production frontiers.

I. INTRODUCTION

The concept of usability, which is concerned with making software systems easy to learn and easy to use, has recently gained increased attention with the development and wide diffusion of end-user interactive software applications (Dray 1995). Usability evaluation, which originally was considered as a burden on software development costs and time, has now taken center stage as an integral part of the software development process. Likewise, various methods for usability evaluation have been proposed and are widely used in practice. The goal of usability evaluation is to identify usability problems in the user interface so that designers can make informed decisions about how to improve the user interface and ultimately achieve a more usable system (Nielsen 1993).

Perhaps, the class of software where the need for achieving usable systems is more pronounced is web-based Internet-based selling websites. There are several reasons why usability of e-commerce websites is more emphasized than with traditional organizational information systems or interactive desktop applications. First, the target users of e-commerce applications are *consumers*, which is untypical of traditional business applications developed for use by employees within a firm (Keeney 1999). As a result, greater constraints are placed on what a designer/developer must do to create a desirable setting for system use by a user/consumer since end-user training is not an option. Second, websites exhibit lower switching costs than do traditional desktop applications or organizational information systems. Since there is no need to purchase (or develop) software prior to installation and use, consumers may easily switch to a competitor's website which is "only a click away".

Current usability evaluation methods can be categorized into the following major classes (Ivory and Hearst 2001):

- *Testing*, which involves an investigator observing users interacting with a user interface to identify usability problems that users run into (e.g., usability testing (Dumas and Redish 1999, Spool et al. 1999)),
- *Inspection*, wherein a usability expert uses a set of criteria (or heuristics) to analyze and critique an interface (e.g., cognitive walkthrough (Wharton et al. 1994), heuristic evaluation (Agarwal and Venkatesh 2002, Nielsen and Molich 1990)),
- *Inquiry*, which engages target users, who are asked to provide feedback on the user interface via structured questionnaires, interviews or focus groups (e.g., QUIS (Questionnaire for User Interface Satisfaction; Chin et al. 1988), the Web Assessment questionnaire (Schubert and Selz 1999)),

- *Analytical modeling*, in which user and interface models are developed to generate predictions about the performance outcomes of the human-computer interaction (e.g., GOMS (Card et al. 1983), CPM-GOMS (John and Kieras 1996)), and
- *Simulation*, in which user models are used to mimic a user interacting with an interface so that simulated data on activities and errors can be measured and reported (e.g., ACT-IF (Pirolli 1997, Pirolli and Card 1999)).

Even though the above approaches to usability evaluation have been successfully applied in the evaluation of user interfaces for traditional organizational information systems and interactive desktop applications, they are not perfectly suited for web-based applications, especially e-commerce applications. First, the iterative development life cycle is dramatically shorter for e-commerce applications. Market pressures for speed often force companies to launch their web-based storefronts prematurely without making sure of the websites' usability. In fact, most managers at e-commerce firms value on-time site launch more importantly than a later launch of a more usable site (Bias 2000). In addition, e-commerce websites are frequently updated and redesigned, which makes the recurring costs of recruiting test users, experts or survey respondents for the evaluation of each redesign excessive for most organizations with limited financial and human resources. Finally, customers display a greater level of heterogeneity of human-computer interaction than users of business applications. This makes it difficult to assume that a large enough set of usability problems will be detected with a limited number of subjects in usability studies (Spool and Schroeder 2001).

Despite these difficulties and challenges, there are opportunities afforded by the Web environment. One notable opportunity stems from the availability of large volumes of data on website usage from log files. Through careful preparation and examination of web server log files, it is possible for usability evaluators to gain insights into how the website is actually being used by the customers. However, the problem with server log data is that there is often too much data available[1]. This situation calls for automated usability evaluation methods for handling the considerable amount of available data (Byrne et al. 1994). In an extensive review of automated methods for usability evaluation, Ivory and Hearst (2001) noted that although methods for automatic *capture* of usability data has been extensively developed, methods for automatic *analysis* (or *critique*) of an interface are still quite limited and are in need of innovative research

[1] A study of by NetGenesis showed that even though e-commerce managers acknowledge that immensely valuable information is contained in the server logs, they are stymied in their desire to access this information due to a lack of people, resources, standard definitions and domain expertise (Cutler and Sterne 2000).

and development. Indeed, given the critical importance of web usability for e-commerce, one of the most pressing questions on the minds of e-commerce managers is whether the design of their online storefronts is effective, and if not, which areas require attention and improvements. Answers to such questions allow managers to prioritize design and redesign projects to maximize return on investment of the firm's development initiatives. This paper seeks to contribute to this stream of research by proposing a usability evaluation method for automatic analysis of e-commerce websites for identifying (potential) usability problems. Our proposed method makes extensive use of actual customer-website interaction data using web server logs. We believe that an automated approach to data collection has the potential to resolve some of the aforementioned problems of current usability evaluation methods: (1) web server logs can be collected continuously and automatically, enabling on-going website evaluations without incurring extraneous costs; (2) data can be collected for all customers making it possible to effectively cope with heterogeneity in consumer behavior.

The paper is organized as follows. The next section presents a production model of online shopping that provides the theoretical and analytical foundation for our empirical method for usability evaluation of e-commerce websites. The third section outlines the empirical method for website evaluation based on this conceptualization. We use DEA (data envelopment analysis) as the methodological vehicle for assessing e-commerce website usability. While doing so, define two metrics, *InefficiencyBreadth* and *UnitInefficiency*, to help identify website functionalities that are less than effective. The *InefficiencyBreadth* metric quantifies the extent of inefficiencies for each website functionality, whereas the *UnitInefficiency* metric computes the severity of inefficiencies for each of the website functionalities. In the fourth section, we illustrate the usefulness of the proposed method by applying it to the evaluation of a real-world e-commerce website. The paper concludes with discussions of the contributions as well as the limitations of the proposed approach. We also discuss several areas of extension for future research.

2. A PRODUCTION MODEL OF ONLINE SHOPPING

Before presenting the details of our method for identifying usability problems in e-commerce websites, we first need to outline the conceptual model that provides the basis for the analytical methods. We conceptualize consumer-website interaction during online shopping as a *production process* in which the customer conducts a purchase transaction by utilizing various

functionalities provided by the e-commerce website[2]. In economics, the production process defines the technical means by which inputs (e.g., materials and resources) are converted into outputs (e.g., goods and services). This technical relationship is represented by the production function, which articulates the maximum level of outputs produced for each given level of inputs (i.e., the efficient frontier or the "best practice" production frontier). Deviations from the production frontier reflect inefficiencies in production (Aigner and Chu 1968). In the context of online shopping, the *inputs* consist of the customers' use of the various functionalities provided by the e-commerce website. They represent the effort put forth by the customer in filling their virtual shopping carts, for example, the number of product page views, extent of navigation through product listings, and references to help pages. The *outputs* of the production process are the contents of the purchase transaction. For example, the number (or dollar value) of items purchased during a shopping trip can be regarded as outputs of the production process. Other factors may additionally impact the efficiency of the production process. For instance, a customer's general competence and skill level with computers and the Internet, her familiarity with a particular e-tailer's website design, and the speed of her Internet connection all could impact how efficient the customer is in producing an online transaction. Borrowing the formalism from production economics, we represent the online shopping service production process as the following *production model* or as its inverse (or "dual"), the *cost model*:

$$\mathbf{y} = f(\mathbf{x}, \mathbf{s}, \boldsymbol{\varepsilon}^{output}) \qquad \textbf{(Production Model)}$$
$$y_r = f(x_i, s_k, \varepsilon_r^{output})$$

or

$$\mathbf{x} = g(\mathbf{y}, \mathbf{s}, \boldsymbol{\varepsilon}^{input}) \qquad \textbf{(Cost Model)}$$
$$x_i = f(y_r, s_k, \varepsilon_i^{input})$$

where

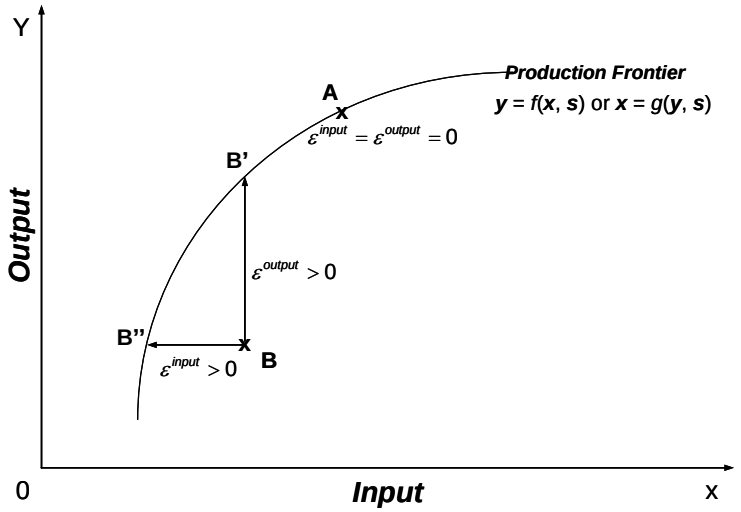
[2] This conceptualization is based on literatures in service operations management (Chase 1978, Chase and Tansik 1984, Lovelock and Young 1979, Mills and Morris 1986, Walley and Amin 1994), and service marketing (Meuter et al. 2000, Zeithaml et al. 1990). See Hahn and Kauffman (2005) for a detailed discussion of the appropriateness and applicability of this conceptualization.

$f(\cdot)$ = production function that translates inputs into outputs,
$g(\cdot)$ = cost function that translates outputs into inputs,
$\mathbf{y}$ = vector of r outputs (y_r) resulting from the production; $r > 0$, $y_r \geq 0$,
$\mathbf{x}$ = vector of i inputs (x_i) used in the production process; $i > 0$, $x_i \geq 0$, $\mathbf{x} \neq 0$,
$\mathbf{s}$ = vector of k environmental variables (s_k) influencing production process,
$\boldsymbol{\varepsilon}^{output}$ = vector of r deviations from production frontier (ε_r^{output}), r > 0, $\varepsilon_r^{output} \geq 0$,
$\boldsymbol{\varepsilon}^{input}$ = vector of i deviation from the production frontier (ε_i^{input}), $i > 0$, $\varepsilon_i^{input}) \geq 0$.

The distinction between output-oriented production model and the input-oriented cost model is useful due to several reasons. First, the different perspectives provide us with flexibility to capture distinctive purchasing behaviors (e.g., goal-directed purchasing vs. hedonic shopping) that have been identified in the marketing literature (Babin et al. 1994, Bloch et al. 1986, Moe and Fader 2001). Goal-directed purchasing typically entails a consumer who has a target product in mind. Hence, her purchasing process is geared toward finding that product with the least amount of effort. For example, a consumer who is shopping for a particular brand of cereal would exhibit goal-directed purchasing behavior. The *input-oriented* cost model, which attributes greater efficiency to production processes which utilize less input given a certain level of output, would be more appropriate for modeling such goal-directed purchasing behaviors. On the other hand, hedonic shopping occurs when a consumer does not have a particular product in mind but is casually browsing through the store to find an item that might catch her fancy (e.g., shopping for pleasure, impulse purchasing). An illustrative example would be a consumer who is shopping for clothes. In this case, the *output-oriented* production model, which attributes greater efficiency to production processes which produce more outputs with a given level of inputs, would be more appropriate for modeling such hedonic shopping behaviors.

Second, this distinction is also useful because it provides an analytical basis for interpreting the inefficiencies in the online shopping behaviors. Inefficiencies in the output-oriented production model (ε_r^{output}) relate to *slack output* – more outputs could have been produced with the same amount of inputs, whereas inefficiencies in the input-oriented cost model (ε_i^{input}) relate to *excess input* – the same amount of outputs could have been produced with less input. Figure 1 shows the basic intuition.

Figure 1. Production Frontiers



The production (cost) function frontier represents the most efficient production (cost) process. All points that lie on the curve (e.g., point A) are said to be *efficient* since they do not deviate from the frontier ($\boldsymbol{\varepsilon}^{output} = \boldsymbol{\varepsilon}^{input} = 0$). All observations that lie below (above) the production (cost) curve (e.g., point B) are *inefficient*. A level of output greater by $\boldsymbol{\varepsilon}^{output}$ may be achieved with the same level of input (i.e., point B') or that the same level of output may be achieved with $\boldsymbol{\varepsilon}^{input}$ less input (i.e., point B").

3. USABILITY EVALUATION: A DEA APPROACH

Conceptualizing online shopping as production enables us to develop a novel perspective for e-commerce website evaluation[3]. Since customers are producing a transaction through the e-commerce website, the e-commerce website can be viewed as a service production environment. The usability of the e-commerce website can thus be assessed by examining how well the production environment (i.e., the e-commerce website) supports efficient transformation of inputs (i.e., user interaction with the website) into outputs (i.e., purchase transactions). Furthermore, we may utilize frontier estimation methods from production econometrics –

[3] The International Organization for Standardization (ISO) defines usability as "… the quality of use – the effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments…" (ISO 1998). Typically, effectiveness is measured by number of errors (or lack thereof), efficiency as task completion time, and satisfaction via self-reported survey instruments. We note that our proposed method cannot capture the satisfaction dimension of usability since we are not making use of primary data collection. Rather, our proposed method focuses on the dimensions of effectiveness and efficiency only.

empirical methods typically used for productivity analysis – in evaluating e-commerce website performance. Of the various analysis methods available, we selected *data envelopment analysis* (DEA), which is a linear programming based non-parametric method for production frontier estimation. We chose it because DEA only requires simple assumptions of monotonically increasing and convex input-output relationships and does not impose strict assumptions with respect to the functional form of the production function. Moreover, DEA can effectively handle production functions where multiple inputs and multiple outputs are involved. Prior research has also shown that the parametric formulation for stochastic frontier estimation and the non-parametric formulation of DEA yield similar results (Banker et al. 1991).

3.1. Data Envelopment Analysis (DEA)

In DEA, the unit of analysis is called the *decision-making unit* (DMU). This represents a production unit. A DMU may be defined narrowly as an individual or as broadly as a firm, an industry, or even as an economy. DEA estimates the relative efficiencies of DMUs from observed measures of inputs and outputs. The productivity of a DMU is evaluated by comparing it against a hypothetical DMU that is constructed as a convex combination of other DMUs in the dataset. Several variants of DEA are available to the analyst to fit the situation at hand. The analyst may choose between input-oriented or output-oriented DEA models. This choice reflects the distinction between the input minimization and the output maximization perspectives, as discussed above. In addition, the analyst may choose between CCR and BCC models depending on whether the production process exhibits constant or variable returns to scale. The CCR model (Charnes et al. 1978, 1981) allows for constant returns to scale, whereas the BCC model (Banker et al. 1984) allows for variable returns to scale. By combining these two considerations, the analyst may model a wide variety of situations. For example, the input-oriented BCC model is appropriate for estimating the productivity of DMUs in terms of input minimization when the production process exhibits variable returns to scale.

The efficiency h_{j0} of DMU j_0, characterized on the basis of the consumption of inputs x_{ij0} and production of outputs y_{rj0}, is assessed by solving the linear program shown in Table 1.

Table 1. DEA Models for Goal-Directed Purchasing and Hedonic Shopping

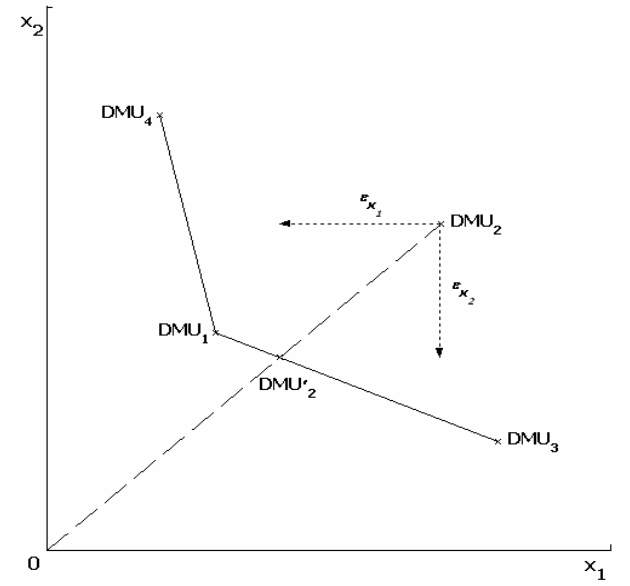
Goal-Directed Purchasing (Input-oriented BCC Model)	Hedonic Shopping (Output-oriented BCC Model)
Min h_{j_0} subject to:	Max h_{j_0} subject to:
$h_{j_0} x_{ij_0} = \sum_{j=1}^{J} x_{ij}\lambda_j + \varepsilon_{ij}^{input}$, $i = 1,\ldots, I$ inputs	$x_{ij_0} = \sum_{j=1}^{J} x_{ij}\lambda_j + \varepsilon_{ij}^{input}$, $i = 1,\ldots,I$ inputs
$y_{rj_0} = \sum_{j=1}^{J} y_{rj}\lambda_j - \varepsilon_{rj}^{output}$, $r = 1, \ldots, R$ outputs	$h_{rj_0} y_{rj_0} = \sum_{j=1}^{J} y_{rj}\lambda_j - \varepsilon_{rj}^{output}$, $r = 1,\ldots,R$ outputs
$\lambda_j \geq 0$, for $\forall j$	$\lambda_j \geq 0$, for $\forall j$
$\sum_{j=1}^{J} \lambda_j = 1, j = 1,\ldots,J$ observations	$\sum_{j=1}^{J} \lambda_j = 1, j = 1,\ldots,J$ observations

The first constraint ensures that all observed input combinations lie on or within the production possibility set defined by the production frontier (i.e., the envelopment condition). The second constraint maintains that the output levels of inefficient observations are compared to the output levels of a reference DMU that is composed of a convex combination of observed outputs. The third constraint ensures that all values of the production convexity weights are greater than or equal to zero so that the hypothetical reference DMU is within the production possibility set. The final constraint allows variable returns to scale[4]. Taken together, the specification of the constraints is such that the production possibilities set conforms to the axioms of production in terms of convexity, monotonicity, variable returns to scale and minimum extrapolation (Banker et al. 1984).

The DEA program is run iteratively for all DMUs ($j = 1,\ldots,J$) to yield efficiency scores h_j^*. A DMU j is said to be fully efficient if the optimal solution h_j^* to its linear program yields $h_j^* = 1$ without any slack output or excess input (i.e., $\varepsilon_r^{output} = \varepsilon_i^{input} = 0$, $\forall i, r$). All other DMUs with $0 \leq h_j^* < 1$ are said to be inefficient (i.e., $\varepsilon_{ij}^{input} > 0$ or $\varepsilon_{rj}^{output} > 0$, $\exists i, r$). The logic behind DEA is shown in Figure 2 with an input-oriented scenario involving two inputs and one output.

[4] This constraint is relaxed for the CCR model that restricts the production process to have constant returns to scale.

Figure 2. Production Frontier and Production Inefficiency in DEA



Note: The graph represents an output isoquant. The inputs, x_1 and x_2, are shown as the axes. All data are normalized for unit output.

The empirical best-practice production frontier is shown by the line segments connecting DMU_4, DMU_1 and DMU_3. Since DMUs 1, 3 and 4 lie on the frontier, they are efficient (i.e., $h^*_{DMU1} = h^*_{DMU3} = h^*_{DMU4} = 1$). DMU_2, however, is inefficient. Compared to the hypothetical DMU'_2 (a convex combination of DMU_1 and DMU_3), the same level of output could have been produced with ε_{x1} and ε_{x2} less inputs. So DMU_2 exhibits excess inputs of ε_{x1} for input x_1 and ε_{x2} for input x_2. The optimal solution h^*_{DMU2} when the DEA model is solved for DMU_2 is the ratio of the distance between the origin and DMU'_2 and that between the origin and DMU_2.

3.2. Identifying Usability Problems

Efficiency estimation via DEA produces efficiency scores, h^*_j, for each transaction. Hence, we may gain an overall assessment of the effectiveness of the e-commerce website by examining the distribution of these efficiency scores or inefficiency deviations, $\theta^*_j = 1/h^*_j - 1$. If most efficiency scores lie close to the efficiency frontier (i.e., $h^*_j \approx 1$ or $\theta^*_j \approx 0$), then we may infer that the e-commerce website is quite effective[5]. However, recall from our earlier discussion that

[5] We note that we are taking a novel approach to using DEA. In prior studies using DEA, the focus was typically on generating efficiency scores and comparing individual DMUs to identify which observations are efficient and which are not. In our paper, we are not particularly interested in the efficiency scores of the DMUs *per se*. Our focus is

an important managerial concern is to understand not only how the e-commerce website is performing, but more importantly, which areas of the website are not effective, so as to identify areas for improvement. In such cases, overall efficiency scores do not help us since the efficiency score relates to the productivity of the production environment as a whole (i.e., the e-commerce website). Instead we would need to delve deeper into the potential causes of the observed inefficiencies. One straightforward way to do this is to examine the breadth (or scope) of observed inefficiencies for each website functionality (i.e., how many transactions exhibited inefficiencies with respect to a particular website functionality) and the severity (or scale) of the observed inefficiencies for each website functionality (i.e., the level of observed inefficiencies when inefficiencies are observed for a particular website functionality). Toward this goal, we define two metrics:

- **Definition 1 (Inefficiency Breadth)**. The *InefficiencyBreadth* of website functionality represents how widespread inefficiencies due to the particular website functionality are.
- **Definition 2 (Unit Inefficiency)**. The *UnitInefficiency* of website functionality on output represents how much the inefficiencies due to the particular website functionality are with respect to a unit of output.

The two metrics above can be easily computed from the DEA results. The *Inefficiency-Breadth* metric can be calculated by determining the proportion of observations for which input inefficiencies were observed: $InefficiencyBreadth_i = n_i / J$. Since input i in the online shopping production model is conceptualized as the customer's use of website functionality i, all non-zero ε_{ij}^{input} represent excess input in the use of website functionality i that resulted in the inefficiency in the production of output r. If we define the set $D_i = \{ j \in J \mid \varepsilon_{ij}^{input} > 0 \}$ (i.e., all DMUs where inefficiency in the use of website functionality i was observed) and $n_i = |D_i|$ (i.e., the cardinality of D_i, the number of elements/observations in set D_i), the proportion of n_i with respect to the total number of DMUs (J) represents the scope of inefficiency due to functionality i.

InefficiencyBreadth is a proportional measure. It basically counts the number of observations (i.e., transactions) that have inefficiency (excess input, ε_{ij}^{input}) observed with respect to a particular input measure (website functionality i) out of the population of all observations (i.e., J). For instance, if we have a total of 100 observations ($J = 100$), and of those, 20 observations

---

on the *distribution* of the efficiency scores. This shift in focus allows us to generate insights about the overall effectiveness of the production environment.

exhibited input inefficiencies for input x_1 (e.g., the search function), then $InefficiencyBreadth_1$ would be 20/100 = 0.20 (or 20%). In other words, this would mean that 20% of all transactions had inefficiencies with respect to the use of the search functionality of the website.

The *UnitInefficiency* metric, which represents the severity of observed inefficiencies for specific website functionality, can be determined by analyzing the magnitude of observed input inefficiencies (ε_{ij}^{input}). Since, observations have differing levels of outputs, we normalize by output: $UnitInefficiency_{irj} = \varepsilon_{ij}^{input} / y_{rj}$. For instance, if the first observation ($j = 1$) had input inefficiency with respect to the first website functionality x_1 (e.g., search) and the actual measure of that inefficiency was 5 (i.e., $\varepsilon_{11}^{input} = 5$). If we further assume that the output (e.g., number of products purchased) of this observation was 50 (i.e., $y_1 = 50$). Then, the $UnitInefficiency_{111}$ is 5/50 = 0.1. In other words, this observation exhibited input inefficiency of 5 (i.e., $\varepsilon_{11}^{input} = 5$) but since her output was quite large (i.e., $y_1 = 50$), on average her inefficiency with respect to input x_1 is 0.1. Note that the *UnitInefficiency* measure is computed for each transaction ($j = 1$ to J observations) for each input x_i ($i = 1$ to I inputs) and for each output y_r ($r = 1$ to R outputs). Hence, we need to investigate distributional measures (e.g., mean, median, variance etc.) in order to interpret the results.

We illustrate the value of our proposed website evaluation method by applying it to a real world operational e-commerce website. Details of the empirical application are presented next.

4. EMPIRICAL APPLICATION

4.1. Research Site and Data Collection

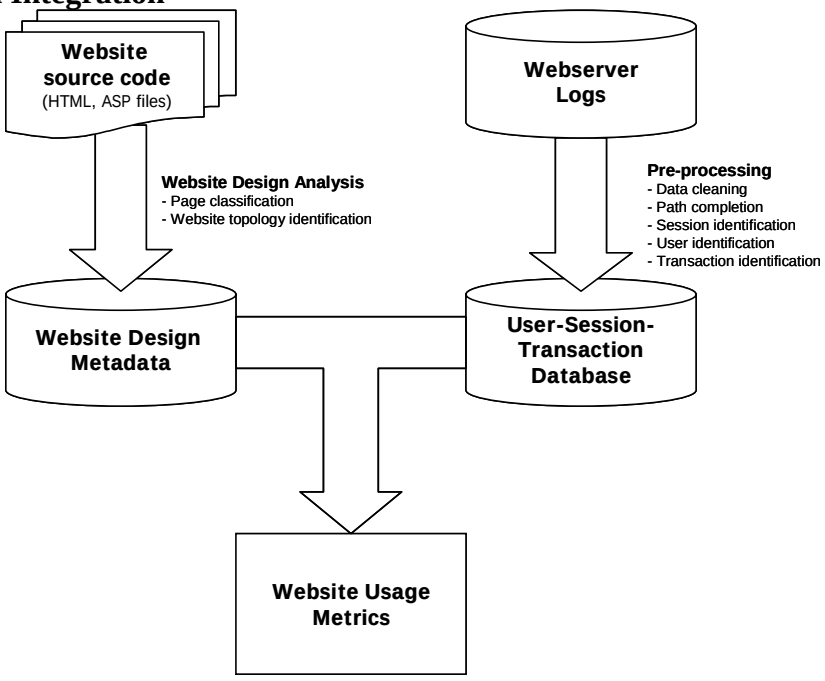
Data for this study were collected at an online grocery retailer, which requires anonymity from the authors. The online grocer is a pure-play Internet-based retailer that delivers groceries directly to its customers' doorsteps with the mission of "taking the dread out of grocery shopping." The company made its first delivery in April 1999. By mid-July 2000, it had over 9000 customers who generated more than $16 million in revenue. At the time of the study, the organization operated only in one metropolitan area in the upper Midwest, where it was the only online grocer in its regional market.

Clickstream data were collected directly from the online grocer's web servers. The website uses HTTP session cookies downloaded onto the visitor's computer to track the customer's shopping behavior at the website. Typical data pre-processing procedures for using web server

logs were used to extract navigation path sequences for visitors from the clickstream data (Cooley et al. 1999). The navigation sessions were combined to identify purchase transactions. Then website usage metrics were extracted to measure the extent to which various areas of the website were used in each of the purchasing processes. The data span two weeks from June 23 to July 5, 2001. A total of 36,051 sessions were recorded for 18,297 unique customers. Our analysis focuses on 5,383 actual completed purchase transactions from 4,941 customers.

We make use of three data sources for analyzing the value of e-commerce website designs: Web server logs, marketing data, and website meta-data. Web servers automatically log all client-server interactions in several log files: access logs, error logs and cookie logs. Error logs store data on failed requests (e.g., missing pages or images and HTTP authentication failures). Although error logs may be used to detect incorrect links or server capacity problems, its use has been proven rather limited for evaluating e-commerce performance. Cookie logs store customized information related to cookies which can be used to identify and track individual users even though they may have used different IP addresses between sessions. Finally, access logs are the most commonly used log files for Web usage mining. Access logs store detailed information about HTTP requests (i.e., date and time of request, client IP address, HTTP auth username, number of bytes transferred, server IP address, the URL stem of request, query string of request, HTTP status, HTTP version, user agent, cookie ID and referrer URL). Web usage mining techniques are performed on the access logs to extract user sessions and patterns of usage behaviors. Next, website meta-data need to be defined to adequately represent the structure and content of the e-commerce website. Website structural meta-data provide the topology of the website (e.g., entry points, the product hierarchy, link structures, etc.) that can be represented as a directed graph structure. Website meta-data also need to provide semantic information related to the content on the individual pages, which can be represented as keyword matrices. Website meta-data are used to make sense of the usage behavior patterns by identifying the nature and content of the pages the customers have gone through to browse and purchase products on the website. The various sources of data can be integrated by joining the heterogeneous data with the appropriate fields. The mapping of the various data sources is presented in Figure 3.

Figure 3. Data Integration



4.2. DEA Model Specification

For our current analyses, we employ the *input-oriented BCC DEA model* (Banker et al. 1984) to estimate the efficiencies of the online purchase transactions in the evaluation of the effectiveness of the online grocer's e-commerce website. Before presenting the specification of the input and output variables, we first discuss the rationale for selecting the input-oriented model (in lieu of the output-oriented model) as well as the BCC model (over the CCR model).

As discussed previously, the input-oriented DEA model with its focus on input minimization (for a given level of output) is appropriate for modeling online purchase situations where goal-directed purchasing is prevalent. Shopping efficiency is meaningful and important in the online grocery shopping context that is investigated here. In fact, the target consumer market for the grocery shopping website is the time-pinched customer who seeks convenience in her grocery shopping activities. Consequently, one of the key operational goals for website design set forth by the managers at the research site is to have first-time customers be able to checkout (i.e., complete a purchase transaction) within an hour and have their subsequent transaction sessions not exceed thirty minutes. Hence, shopping efficiency is indeed a major focus in the current context.

We also employ the BCC model (Banker et al. 1984) because it allows for variable returns to scale in the production process. The CCR model (Charnes et al. 1978, 1981) enforces constant returns to scale in the production process, and so it is less desirable for this context. This is because in the online shopping context the size of the transaction (i.e., the number of items purchased), as an indicator of scale size, is under the control of the customer and not the e-commerce firm. Hence, even if an optimal production scale were to exist and be found, one cannot enforce it[6]. Besides the theoretical reason for this choice, it is also possible to determine empirically whether the production function exhibits constant or variable returns to scale (Banker and Slaughter 1997). Our analyses show that the online shopping production process does in fact exhibit variable returns to scale, hence the BCC model formulation is appropriate here.

Finally, DEA analysis is only as good as the initial selection of input and output variables. The inputs must represent the resources consumed by the DMUs and the outputs must characterize the end results of the production by the DMUs. Another conceptualization of the outputs is *unit performance*. This works so long as DEA's axioms of production are satisfied. [7] In online shopping, *inputs* consist of customers' use of various website functionalities and the *output* consists of a checkout of a basket of products. The input and output variables are summarized in Table 2.

[6] Enforcing a production scale on a consumer would be analogous to suggesting that she purchase more (or less) items because she would be more scale efficient with a different basket size.

[7] The *production axioms* form the basic theoretical logic of how production occurs under the *theory of production* in economics. They reflect simple ideas, and structure productivity analysis. For example, the regularity axiom states that you cannot produce something from nothing. Another axiom, the axiom of monotonicity (or inefficiency), is that having more of an input assures you that you are able to produce no less an output level than what you could produce with less of the output. The convexity axiom suggests that if it is possible to produce at two different levels of output, based on some configuration of inputs, then it is possible to achieve convex combinations (or averages) of the output levels with adjustments to the inputs. The ray unboundedness axiom relates to the assumption used to determine constant or increasing returns to scale. Finally, the axiom of minimum extrapolation ensures that production relationships should not be inferred outside the observed production scale.

Table 2. DEA Model's Input and Output Variables

Category	Variable	Measure	Description
Inputs	x_1	*Products*	Number of product page views
	x_2	*Lists*	Number of product list views
	x_3	*Personal*	Number of personal list views
	x_4	*OrderHistory*	Number of order history page views
	x_5	*Search*	Number of search conducted
	x_6	*Promotion*	Number of promotional page views
	x_7	*Recipe*	Number of recipe page views
	x_8	*Checkout*	Number of checkout pages
	x_9	*Help*	Number of help page views
Output	y_1	*BasketSize*	Number of different products at checkout

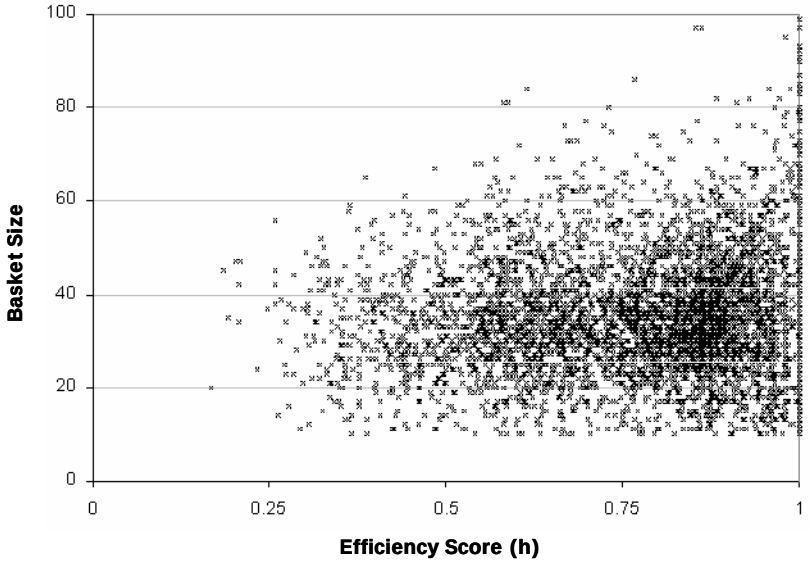
Taken together, the nine input variables (x_1 through x_9) represent all major website functionalities a customer has used in conducting her purchase transaction. The output measure, the number of different products at checkout (y_1 = *BasketSize*), represents the level of performance of the online shopping production process.

4.3. Results

4.3.1. DEA Results.

Figure 3 shows the efficiency scores of all DMUs (J = 5383) against the respective output of each observation. The horizontal axis represents the efficiency scores of the online shopping transactions ($0 \leq h_j^* < 1$), whereas the output level (i.e., number of items in the cart at checkout) is represented on the vertical axis. The efficient transactions lie on (or near) the right edge of the graph ($h^*_j \approx 1$). Visual inspection gives a summary of overall website efficiency. The plot shows significant variability of efficiency scores at all output levels, suggesting that the website may not be entirely effective.

Figure 4. DEA Scores by Output Level



4.3.2. Inefficiency Results by Website Functionality.

To gain insights into the potential causes of the observed overall website inefficiency, we analyzed the inefficiencies by website functionality with the inefficiency metrics proposed earlier. (See Table 3).

Recall that $InefficiencyBreadth_i$ (4th column) measures the proportion of DMUs for which excess input for website functionality i was observed. For example, of all purchase transactions ($J = 5383$), excess input for the first website functionality (*ProductInformation*, DEA input variable x_1) was observed for 2272 DMUs ($n_1 = 2272$), $InefficiencyBreadth_1$ for *ProductInformation* is 42.21% (i.e., 2272/5383 = 0.4221). We see that of the various website functionalities, *InefficiencyBreadth* was greatest for *ProductInformation* (42.21%), then *Promotion* (39.4%), *PersonalList* (31.4%) and *Search* (25.6%). With the remaining five website functionalities (i.e., *ProductList*, *OrderHistory*, *Recipe*, *Checkout* and *Help*), the breadth of inefficiency was less significant.

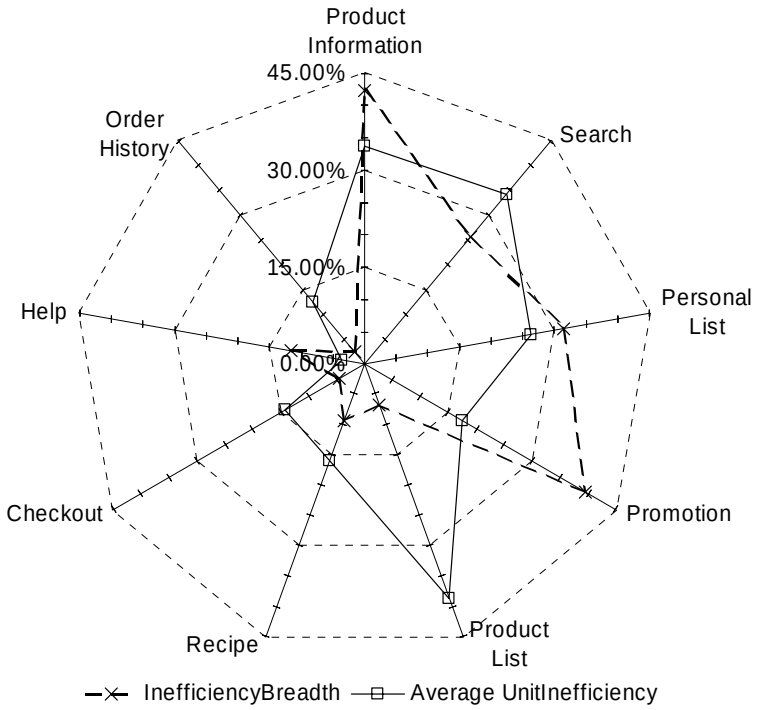
Table 3. Inefficiency of Website Functionality

Variable (*i*)	Website Functionality	n_i	Inefficiency Breadth	UnitInefficiency Average	UnitInefficiency Total
x_1	*ProductInformation*	2272	42.21%	0.3368	765.27
x_2	*ProductList*	359	6.67%	0.3851	138.24
x_3	*PersonalList*	1690	31.40%	0.2609	440.86
x_4	*OrderHistory*	128	2.38%	0.1248	15.97
x_5	*Search*	1378	25.60%	0.3415	470.60
x_6	*Promotion*	2121	39.40%	0.1763	373.88
x_7	*Recipe*	499	9.27%	0.1580	78.83
x_8	*Checkout*	240	4.46%	0.1427	34.25
x_9	*Help*	621	11.54%	0.0348	21.63

Note: The total in the right-hand column is the sum of all *UnitInefficiency* values for all *j* DMUs.

Even though *InefficiencyBreadth* provides a useful metric that represents how widespread inefficiencies due to specific website functionality were, it does not provide much information as to the severity of those observed inefficiencies. Thus we also investigate the *UnitInefficiency* metric to gain more insights. Figure 5 shows a radar chart for these metrics.

Figure 5. *InefficiencyBreadth* and *UnitInefficiency* by Website Functionality



The chart includes *InefficiencyBreadth* (dashes) and average *UnitInefficiency* (solid line) sorted by decreasing order of total *UnitInefficiency* (clockwise starting from the top). The results show that website functionalities *ProductInformation*, *Search* and *PersonalList* were most

problematic in incurring inefficiencies at the e-tailer's website. Inefficiencies due to website functionalities *OrderHistory*, *Help*, *Checkout* and *Recipe* also were quite insignificant. For example, we see that the website functionality of *ProductInformation* was the area in which inefficiencies were not only the most widespread (*InefficiencyBreadth* = 42.21%) but also the most severe (average *UnitInefficiency* = 0.337).

On the other hand, inefficiencies due to *OrderHistory* were neither widespread nor serious (*InefficiencyBreadth* = 2.38%, average *UnitInefficiency* = 0.125). We also observe interesting results for website functionalities, *Promotion* and *ProductList*. Inefficiencies due to *Promotion* were widespread (*InefficiencyBreadth* = 39.4%), but the degree of inefficiency was low (average *UnitInefficiency* = 0.176). Meanwhile, the scope of inefficiencies due to *ProductList* was narrow (*InefficiencyBreadth* = 6.67%) but the degree of inefficiency was quite substantial (average *UnitInefficiency* = 0.385).

We may also formally test the differences in the *UnitInefficiency* scores between the website functionalities to gain more confidence in the interpretation of the results. In order to formally test differences in efficiencies, the statistical test procedure proposed by Banker (1993) can be used for comparing efficiency ratings between groupings. Basically, the statistical procedure involves testing whether the means of the inefficiency score probability distributions for different conditions are different. Two test statistics were proposed by Banker depending on whether inefficiency deviations of the observed data are postulated to be drawn from an *exponential* or a *half-normal distribution*[8].

The overall test procedure is as follows. Let *j* represent an online shopping transaction in the overall dataset. The *UnitInefficiency* score of a shopping transaction *j* in group D_i is denoted by $\theta_j^{D_i}$. If one assumes the inefficiency deviations to be exponentially distributed with parameter σ_i, the null hypothesis for comparing two groups pair-wise (i.e., *UnitInefficiency* scores for two website functionalities, say *a* and *b*) is that inefficiencies due to the two website functionalities are not different, H_0: $\sigma_a = \sigma_b$. The alternative hypothesis is H_1: $\sigma_a > \sigma_b$: the inefficiency level due to website functionality *a* is greater than those due to website functionality *b* (i.e., website functionality *a* is showing more inefficiencies than website functionality *b*). The test statistic is:

[8] It is reasonable to assume an exponential distribution for the inefficiency deviations when one has reason to believe that most observations are close to the production frontier, whereas a half-normal distribution should be assumed when few observations are likely to be are close to the frontier.

$$\frac{\sum_{j \in D_a} (\theta_j^{D_a} - 1)/n_a}{\sum_{j \in D_b} (\theta_j^{D_b} - 1)/n_b}$$

The test statistic asymptotically follows the F-distribution with $(2n_a, 2n_b)$ degrees of freedom for large n, where n_a and n_b are the number of observations in the subsets D_a and D_b, respectively. On the other hand, if one assumes the inefficiency deviations to be half-normally distributed then a different test statistic is used:

$$\frac{\sum_{j \in D_a} (\theta_j^{D_a} - 1)^2/n_a}{\sum_{j \in D_b} (\theta_j^{D_b} - 1)^2/n_b}$$

This statistic again asymptotically follows an F-distribution with (n_a, n_b) degrees of freedom for large values of n.

We conducted pair-wise comparison of the *UnitInefficiency* scores for each of the website functionalities. The results are summarized in Table 4.

Table 4. Statistical Pair-wise Comparison of UnitInefficiency Scores

Website Functionality Dimensions	**Website Functionality Dimensions**								
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
(1) *ProductList*		H	E	E, H	E, H	E, H	E, H	E, H	E, H
(2) *Search*				E, H	E, H	E, H	E, H	E, H	E, H
(3) *ProductInformation*				E, H	E, H	E, H	E, H	E, H	E, H
(4) *PersonalList*					E, H	E, H	E	E, H	E, H
(5) *Promotion*							E, H	E, H	E, H
(6) *Recipe*								H	E, H
(7) *Checkout*					H			H	E, H
(8) *OrderHistory*									E, H
(9) *Help*									

Note: The comparisons are from *row* to *column*. "E" denotes *statistically significant differences* in *UnitInefficiency* scores between the website functionality of the row and the website functionality of the column under the assumption of *exponentially-distributed* *UnitInefficiency* scores. "H" denotes a statistically significant difference when assuming the *UnitInefficiency* scores follow a *half-normal distribution*.

We conducted all of the statistical tests with significance levels at $\alpha = 0.01$. The results show that, with a few exceptions, the rank ordering of the severity of inefficiencies by website functionality seems to represent quite distinct levels of severity. For example, we see that inefficiencies due to *ProductList* are more severe (in the statistical sense) than all inefficiencies

due to all other website functionalities. Inefficiencies due to *Search* and *ProductInformation* are similar (i.e., not statistically different) but these inefficiencies are more severe than inefficiencies due to all other website functionalities except *ProductList* (i.e., *PersonalList*, *Promotion*, *Recipe*, *Checkout*, *OrderHistory* and *Help*). The remainder of the results table can be interpreted in a similar manner[9].

Until now, we have presented general results from using the website evaluation method in identifying potentially problematic website areas. We note that there are numerous other ways in which insightful analyses can be conducted. A simple extension would be to divide the dataset into multiple groups to see if observed website inefficiencies are similar (or different) across groups of customers. For example, the dataset could be divided based on length of relationship with the online service (e.g., loyal and frequently returning customers vs. newly registered customers) to see whether or not these different groups exhibit differences in inefficiencies or if a different set of website functionalities are problematic for different groups.

4.3.3. Inefficiency Results by Output Volume.

To further demonstrate the value of the proposed evaluation method, we present additional results from a simple extension – investigating website inefficiencies by output volume. The key question that guides this analysis is whether output volume (i.e., cart size) has an impact on efficiency. In other words, since customers that conduct high volume transactions may exhibit different purchasing and website interaction behaviors from those that conduct lower volume ones, we analyzed the DEA results to explore these issues.

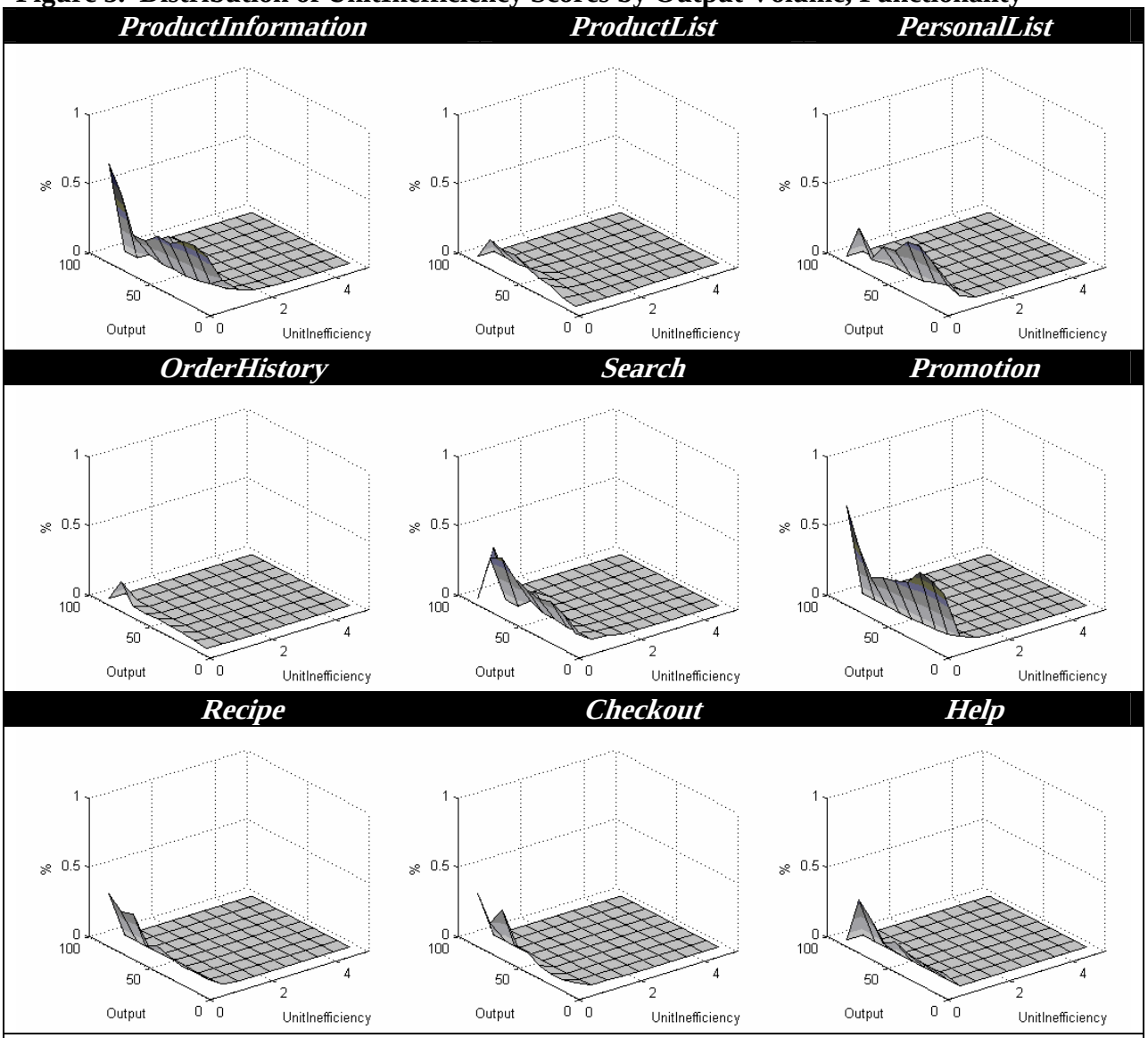
Figure 5 shows the distributions of *UnitInefficiency* values by output volume for each of the website functionalities. Several results are noteworthy. The distribution of *UnitInefficiency* values is skewed with most observations in the lower range (with a long tail). *UnitInefficiency* seems to follow an exponential or half-normal distribution rather than a symmetric distribution such as the normal distribution. Second, we reconfirm some of the insights generated previously. On average, *UnitInefficiency* was most salient for website functionalities *ProductInformation*,

---

[9] Note that we are naively presenting these statistical test results without making use of additional background information that designers would typically have in practice. For example, although we find that the website functionalities of Product Information, Search and Personal List are the most problematic, the designers of the website may not find this result particularly alarming if they have reason to believe that those functionalities are inherently inefficient. In other words, website designers may (and should) incorporate their prior knowledge as benchmark expectations. The purpose of this exposition is to illustrate that such statistical testing is possible, not to formally test any hypothesis about the efficacy of various website functionalities.

Promotion, *Search* and *PersonalList*. This can be seen by observing the height of the surface plots. A closer look at the results indicates that higher volume transactions seem to be relatively more likely to incur inefficiencies than lower volume ones, regardless of the website functionality. This suggests that the website may be geared toward supporting smaller carts.

Figure 5. Distribution of UnitInefficiency Scores by Output Volume, Functionality



Note: The height of the surface shows the proportion of transactions in which a particular *UnitInefficiency* value was observed for a particular output volume.

The last finding is an interesting result when we consider the nature of the products being sold by the online grocer and how this impacts consumer purchase behaviors. The overall design

strategy of the current website is typical in Internet-based selling, with hierarchical product categories for drill-down, a search function, basic checkout functions, help pages and promotional pages for specials and recipes. What the results may be suggesting is that even though such a design strategy may be effective for e-tailers where the typical number of products being purchased is small (i.e., cart size of 1 to 5 items), a different overall design strategy may be required for grocery shopping where the number of different items being purchased is larger (i.e., cart size of 40+ items).

5. CONCLUSION

Evaluating the effectiveness of e-commerce website design is an important, yet complex problem for e-commerce retailers. Their success hinges largely on their ability to provide a high-quality website. So e-commerce retailers need to constantly monitor the effectiveness of their web-based storefronts to identify those website areas that are problematic. However, current methods do not offer any practical means for a solution to this problem. We proposed an innovative method for automatically identifying e-commerce website inefficiencies.

By modeling online shopping as an economic production process and using evaluative methods for frontier analysis, we defined and estimated a value-driven model for website effectiveness that allows us to identify problematic areas within the e-commerce website. We also demonstrated the value of our method by applying it to the evaluation of a real-world e-commerce website. Through the application, it was possible to gain a deeper understanding of which website functionalities were potentially problematic. It was also possible to discover unexpected knowledge related to the potential inappropriateness of the overall design strategy of the e-tailer's website. Although we do not have conclusive results with respect to this last point, such knowledge discovery provides a useful starting point for delving deeper into these issues.

The proposed website evaluation method provides significant benefits over current methods that are used widely. The empirical insights generated could not have been uncovered using the traditional methods of user testing, inspection or inquiry. One of the major advantages of the proposed method is that firms can make use of observable customer actions for all users and customers at a given website. In fact, the problem of scalability is a major concern with the previous evaluation methods. With user testing, for instance, deciding on the adequate number of subjects to test for a representative picture of website usability problems is still in debate

(Bevan et al. 2003, Spool and Schroeder 2001). Also, it is difficult for usability experts to be able to identify all usability problems that may arise for the wide variety of different users who may be customers at the website due to bounded rationality (Fu et al. 2002). We are not arguing, however, that traditional testing, inquiry and inspection do not provide value. Instead, these methods have their own specific complementary strengths (especially during the design stages in the systems development life cycle before a site is launched) and should be employed in conjunction with the proposed method. For example, user satisfaction is an important attribute of usability that can only be measured with inquiry techniques (e.g., surveys or interviews).

Our method also provides the benefit of an unobtrusive approach to data collection. Although online user surveys leverage available web technologies, and are widely adopted, response bias (Schwarz 1999) and non-response bias (Andrews et al. 2003) will persist. Moreover, with the frequent website redesigns, it becomes difficult to solicit continuous responses for each redesign. A major benefit of the proposed method is that one may bypass the aforementioned problems by making use of automatically-collected web server logs of customer web navigation behavior that occur in a genuine real world setting. The empirical methods we used do not employ any proprietary data specific to our particular research site. Indeed, we expect that the required data will be available to all e-commerce firms. So the proposed method should be readily applicable to any transactional e-commerce website. With some additional effort, the data preparation and even the DEA programs can be systematically programmed into the web application servers. This makes it possible to automatically generate metrics so that e-commerce managers can continuously *monitor* the effectiveness of their website without incurring the costs of extraneous data collection and tedious analysis.

We should also acknowledge a number of caveats and considerations with respect to the interpretation of the results of this study, as well as the implementation of the proposed method. Even though the value of the proposed website evaluation method can be inferred by the interesting results enumerated above, care must be taken not only when interpreting the results but also when trying to apply the method more broadly. For example, the results show that the some website functionalities (e.g., Product List, Search and Product Information) were more problematic than others (e.g., Help, Order History and Checkout). However, the reader should not over-generalize and assume that these website functionalities would also be problematic on other e-commerce websites. The main focus of our evaluation method proposed here is not in

uncovering universal design guidelines that may be applied to any setting (e.g., identifying the optimal organization of product hierarchies in an e-commerce website). Instead, the focus of the proposed evaluation method is to provide to managers at e-commerce firms with useful feedback concerning how their customers are performing in the presence of their current website designs. As briefly described in the Introduction, the proposed evaluation method is intended to be used as a tool for the continuous management of website quality. The rationale is similar in spirit to an important research stream in software engineering economics, where metrics for evaluating software development and maintenance productivity have been developed as a vehicle for managing and maximizing the value of software development projects (e.g., Banker et al. 1991, Banker and Kauffman 1991, Banker and Slaughter 1997, Chidamber et al. 1998). Likewise, the proposed evaluation method is intended for use within a firm for managing its website development initiatives.

We also note that the data collection window of two weeks could have introduced bias in the dataset as only transactions completed within the two-week window are included. In other words, transactions that take longer than two weeks were dropped from the analysis. This is potentially an important concern since many online consumers engage in "look-to-book" type shopping[10]. Consequently, the results of the current analyses need to be interpreted with this limitation in mind. That said, the proposed method could be applied with a more complex model that tracked consumer purchases for a longer period of time so that such situations could also be handled.

Another limitation stems from the assumptions of the production model. The production model is essentially additive in that the use (or consumption) of inputs contributes *independently* to the production of outputs. In other words, our method currently does not allow for investigating interaction effects in website functionalities. Our future research agenda includes extending the online shopping model so that such interaction effects can also be dealt with.

A final area of potential concern relates to the applicability of our proposed method to a broader context of usability evaluation. In our currently study, we have applied our method to gain insights into the effectiveness of the website of an online grocer. An important

[10] "Look-to-book" type shopping is where the customer adds items to the cart not for the purpose of immediate purchase, but to keep track of items of interest. For example, when a consumer identifies an interesting book on Amazon.com she would put that item in her cart to keep track of it. However, the actual purchase transaction (i.e., checkout) may occur at a later point in time when several such "look-to-book" sessions have been aggregated.

characteristic of the grocery domain is that purchase behaviors are primarily goal-directed. In our analyses, we have instantiated our online shopping model to specifically take into account this aspect. For example, we have modeled the online shopping production process with an *input*-oriented production framework, which is more appropriate for goal-directed purchasing. Other e-commerce websites that deal with different types of goods (e.g., books, CDs, DVDs, apparel etc.) can and should be evaluated with a different modeling perspective depending on the nature of the purchase behaviors that are typically expected for such websites. For example, in the case of Internet-based selling of clothes, consumer purchase behaviors will typically entail experiential (or hedonic) shopping motives. In such cases, the online shopping production process should be instantiated with an *output*-oriented model which focuses on maximizing the level of outputs given a level of inputs. In other words, website navigation behaviors that results in more products identified, considered and purchased (i.e., output) given the same amount of website use (i.e., input) would be regarded as more effective. The proposed model and method is general enough so that these different types of consumer behaviors can be appropriately captured.

Also, we need to think about whether the proposed usability evaluation method can effectively be applied to non-commercial website applications (e.g., search engines, online portals, informational websites etc.) We believe that our proposed method works as long as the production analogy is appropriate. In other words, if we can effectively model the human-computer interaction behavior as a production process (i.e., consumption of inputs in order to produce outputs), then the proposed method should be readily applicable. The production framework and the ensuing efficiency orientation is consistent with information foraging theory (Pirolli and Card 1999), which has been widely used in the HCI literature. Information foraging theory posits that people, whenever possible, modify their information seeking strategies or the structure of the information environment to maximize their rate of gaining valuable information. This is in essence referring to the maximization of outputs (i.e., valuable information) while minimizing inputs (i.e., information seeking strategies), which is in line with the production efficiency paradigm.

We acknowledge, however, that the production framework applies more easily in the context of transactional e-commerce websites due to the relatively unambiguous characterization of inputs, and more importantly outputs (e.g., products purchased). However, for informational websites, for instance, it may be more difficult to distinguish between valid and invalid outputs.

For example, if a user retrieves a particular web page, how can we be sure that that page contains information that the user is seeking, rather than a page a user mistakenly retrieved. In e-commerce, this is less problematic because if a user purchases an item (or adds the item to the virtual shopping cart), then we can reasonably assume that the user was seeking that item. Techniques for inferring user needs from observable user behaviors (e.g., Chi et al. 2001) should prove valuable in these regards.

REFERENCES

Agarwal, R. and V. Venkatesh. 2002. Assessing a firm's web presence: A heuristic evaluation procedure for the measurement of usability. *Information Systems Research* **13**(2) 168-186.

Aigner, D. J. and S. F. Chu. 1968. On estimating the industry production function. *American Economic Review* **58**(4) 826-839.

Andrews, D., B. Nonnecke and J. Preece. 2003. Electronic survey methodology: A case study in reaching hard-to-involve internet users. *International Journal of Human-Computer Interaction* **16**(2) 185-210.

Babin, B. J., W. R. Darden and M. Griffin. 1994. Work and/or fun: Measuring hedonic and utilitarian shopping value. *Journal of Consumer Research* **20**(4) 644-656.

Banker, R. D. 1993. Maximum likelihood, consistency and data envelopment analysis: A statistical foundation. *Management Science* **39**(10) 1265-1273.

Banker, R. D., A. Charnes and W. W. Cooper. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* **30**(9) 1078-1092.

Banker, R. D., S. Datar, M. and C. F. Kemerer. 1991. A model to evaluate variables impacting the productivity of software maintenance projects. *Management Science* **37**(1) 1-18.

Banker, R. D. and R. J. Kauffman. 1991. Reuse and productivity: An empirical study of integrated computer-aided software engineering (icase) at the first boston corporation. *MIS Quarterly* **15**(3) 374-401.

Banker, R. D. and S. A. Slaughter. 1997. A field study of scale economies in software maintenance. *Management Science* **43**(12) 1709-1725.

Bevan, N., C. Barnum, G. Cockton, J. Nielsen, J. M. Spool and D. Wixon. 2003. Panel: The "magic number 5:" is it enough for web testing? *Proceedings of the 2003 ACM Conference on Human Factors in Computing Systems*, Ft. Lauderdale, FL, ACM Press, New York, NY, 698-699.

Bloch, P. H., D. L. Sherrell and N. M. Ridgway. 1986. Consumer search: An extended framework. *Journal of Consumer Research* **13**(1) 119-126.

Byrne, M. D., S. D. Wood, P. N. Sukaviriya, J. D. Foley and D. E. Kieras. 1994. Automating interface evaluation. *Proceedings of the 1994 ACM Conference on Human Factors in Computing Systems*, Boston, MA, ACM Press, New York, 232-237.

Card, S. K., T. P. Moran and A. Newell. 1983. *The psychology of human computer interaction*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Charnes, A., W. W. Cooper and E. Rhodes. 1978. Measuring efficiency of decision-making units. *European Journal of Operational Research* **2**(6) 428-449.

Charnes, A., W. W. Cooper and E. Rhodes. 1981. Evaluating program and managerial efficiency: An application of data envelopment analysis to program follow through. *Management Science* **27**(6) 668-697.

Chase, R. B. 1978. Where does the customer fit in a service operation? *Harvard Business Review* **56**(6) 138-139.

Chase, R. B. and D. A. Tansik. 1984. The customer contact model for organization design. *Management Science* **29**(9) 1037-1050.

Chi, E. H., P. L. T. Pirolli, K. Chen and J. E. Pitkow. 2001. Using information scent to model user information needs and actions on the web. *Proceedings of the 2001 ACM Conference on Human Factors in Computing Systems*, Seattle, WA, ACM Press, 490-497.

Chidamber, S. R., D. P. Darcy and C. F. Kemerer. 1998. Managerial use of metrics for object oriented software: An exploratory analysis. *IEEE Transactions on Software Engineering* **24**(8) 629-639.

Chin, J. P., V. A. Diehl and K. L. Norman. 1988. Development of an instrument measuring user satisfaction of the human-computer interface. *Proceedings of the 1988 ACM Conference on Human Factors in Computing Systems*, Washington, D.C., 213-218.

Cooley, R., B. Mobasher and J. Srivastava. 1999. Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems* **1**(1) 5-32.

Dray, S. M. 1995. The importance of designing usable systems. *Interactions* **2**(1) 17-20.

Dumas, J. S. and J. C. Redish. 1999. *A practical guide to usability testing* (revised ed.). Intellect Books, Exeter, UK.

Fu, L., G. Salvendy and L. Turley. 2002. Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour and Information Technology* **21**(2) 137-143.

Hahn, J. and R. J. Kauffman. (2005). Measuring the effectiveness of e-commerce website design (Working Paper), Management Information Systems Research Center, University of Minnesota. Minneapolis, MN 2005.

ISO. (1998). Ergonomic requirements for office work with visual display terminals (vdts) -- part 11: Guidance on usability (iso 9241-11:1998): International Organization for Standardization.

Ivory, M. Y. and M. A. Hearst. 2001. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys* **33**(4) 470-516.

John, B. E. and D. E. Kieras. 1996. The goms family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction* **3**(4) 320-351.

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

Lovelock, C. H. and R. F. Young. 1979. Look to consumers to increase productivity. *Harvard Business Review* **57**(3) 168-178.

Meuter, M. L., A. L. Ostrom, R. I. Roundtree and M. J. Bitner. 2000. Self-service technologies: Understanding customer satisfaction with technology based service encounters. *Journal of Marketing* **64**(3) 50-64.

Mills, P. K. and J. H. Morris. 1986. Clients as "partial" employees of service organizations: Role development in client participation. *Academy of Management Review* **11**(4) 726-735.

Moe, W. W. and P. S. Fader. 2001. Uncovering patterns in cybershopping. *California Management Review* **43**(4) 106-117.

Nielsen, J. 1993. *Usability engineering*. Morgan Kaufmann Publishers, San Francisco, CA.

Nielsen, J. and R. Molich. 1990. Heuristic evaluation of user interfaces. *Proceedings of the 1990 ACM Conference on Human Factors in Computing Systems*, Seattle, WA, 249-256.

Pirolli, P. L. T. 1997. Computational models of information scent-following in a very large browsable text collection. *Proceedings of the 1997 ACM Conference on Human Factors in Computing Systems*, Atlanta, GA, ACM Press, New York, NY, 3-10.

Pirolli, P. L. T. and S. K. Card. 1999. Information foraging. *Psychological Review* **106**(4) 643-675.

Schubert, P. and D. Selz. 1999. Web assessment: Measuring the effectiveness of electronic commerce sites going beyond traditional marketing paradigms. *Proceedings of the 32nd Hawaii International Conference on System Sciences*, Maui, HI, IEEE Computer Society Press, Los Alamitos, CA.

Schwarz, N. 1999. Self-reports: How the questions shape the answers. *American Psychologist* **54**(2) 93-105.

Spool, J. M., T. Scanlon, W. Schroeder, C. Synder and T. DeAngelo. 1999. *Web site usability: A designer's guide*. Morgan Kaufmann Publishers, San Francisco, CA.

Spool, J. M. and W. Schroeder. 2001. Testing web sites: Five users is nowhere near enough. *Proceedings of the 2001 ACM Conference on Human Factors in Computing Systems*, Seattle, WA, ACM Press, New York, NY., 285-286.

Walley, P. and V. Amin. 1994. Automation in a customer contact environment. *International Journal of Operations and Production Management* **14**(5) 86-100.

Wharton, C., J. Rieman, C. Lewis and P. G. Polson. 1994. The cognitive walkthrough method: A practitioner's guide. *Usability inspection methods*. J. Nielsen and R. L. Mack, eds. Wiley, New York, 105-140.

Zeithaml, V. A., A. Parasuraman and L. L. Berry. 1990. *Delivering quality service: Balancing customer perceptions and expectations*. Free Press, New York, NY.