A DESIGN SCIENCE APPROACH FOR IDENTIFYING USABILITY PROBLEMS IN WEB SITES THAT SUPPORT INTERNET-BASED SELLING

ABSTRACT

Managers at e-commerce firms are in need of proven methods for evaluating the usability of their Web sites. 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 and inquiry) are not well suited to the tasks at hand. This paper proposes a new Web site evaluation approach, which is grounded in the *economic theory of production*. We conceptualize human-computer interaction (HCI) during online shopping as an economic production process in which customers make use of various functionalities provided by the e-commerce Web site to complete a purchase transaction. This view enables us to formulate a novel perspective on Web site usability — the ability to transform inputs (i.e., use of Web site functionalities) into outputs (i.e., completed purchase transactions). We use *data envelopment analysis* (DEA) methods for evaluating e-commerce Web sites, and propose two new metrics, *InefficiencyBreadth* and *UnitInefficiency*, to help identify potentially problematic Web site functionalities. The value of the proposed method is illustrated by applying it to the evaluation of a real-world e-commerce Web site.

Keywords: Design science research, data envelopment analysis; e-commerce; Internet-based selling; performance measurement; software metrics; usability assessment; Web site evaluation.

A DESIGN SCIENCE APPROACH FOR IDENTIFYING USABILITY PROBLEMS IN WEB SITES THAT SUPPORT INTERNET-BASED SELLING

1. INTRODUCTION

In the electronic commerce context, the class of software applications where the importance of understanding end user needs and requirements is most pronounced is Web sites for Internet-based selling. There are several reasons why the usability of Internet-based selling Web sites is emphasized more than traditional organizational information systems (IS) 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 or developer must do to create a desirable setting for system use by a user/consumer since end-user training is not an option (Albert et al. 2004). Second, Web sites exhibit lower search and switching costs than do traditional desktop applications or organizational systems (Bakos 2001; Chen and Hitt 2002). Since there is no need to purchase and install software prior to use, consumers may easily switch to a competitor's Web site which is "only a click away" (Johnson et al. 2003).

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 an added burden on software development costs and time, has now taken center stage as an integral part of the software development process (Anderson et al. 2001). 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). Although a variety of usability evaluation methods have been proposed and are widely used, most can be categorized into the following major classes (Ivory and Hearst 2001):

- *Testing* involves an investigator observing users interacting with a user interface to identify usability problems that users run into. A common example would be usability testing in controlled laboratory settings (e.g., Dumas and Redish 1999, Spool et al. 1999).
- *Inspection* methods involve a usability expert who uses a set of criteria (or heuristics) to analyze and critique an interface. Examples include cognitive walkthroughs (Wharton et al. 1994) and heuristic evaluation (Agarwal and Venkatesh 2002; Nielsen and Molich 1990).
- *Inquiry* methods engage target users, who are asked to provide feedback on the user interface via structured questionnaires, interviews or focus groups. The Questionnaire for User Interface Satisfaction (QUIS) (Chin et al. 1988), the Web Assessment Questionnaire (Schubert and Selz 1999), and facilitated group techniques (Duggan 2003) are well-known examples.
- *Analytical modeling* involves the development of formal models of users and interfaces, which are used to generate predictions about the performance outcomes of HCI. Examples include GOMS (Card et al. 1983) and CPM-GOMS (John and Kieras 1996).

- *Simulation* builds formal user models that mimic a user interacting with an interface so that simulated data on activities and errors can be measured and reported. ACT-IF (Pirolli 1997; Pirolli and Card 1999) is a notable case of the simulation method.

Even though these approaches to usability evaluation have been successfully applied in the evaluation of user interfaces for traditional organizational IS 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 (Gallaugher 1999). Market pressures for speed often force companies to launch their Web-based storefronts prematurely without making sure of the Web sites' usability. In fact, most managers at e-commerce firms value on-time site launch more than a later launch of a more usable site (Bias 2000). In addition, e-commerce Web sites are frequently updated and redesigned (Brajnik 2000), which makes the recurring costs of recruiting test users, experts or survey respondents, and/or the development of user and interface models for the evaluation of each redesign excessive for most organizations with limited financial and human resources. Finally, general consumers display a greater level of heterogeneity of human-computer interaction (HCI) than managers or users of business applications (Albert et al. 2004). 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 clickstream data on Web site 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 Web site is actually being used by the customers (Drott 1998; Fuller and de Graaff 1996; Rosenstein

2000). 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 and development.

We seek to contribute to this stream of research by conducting a design science research that develops a usability evaluation method for *automated* analysis of e-commerce Web sites for identifying potential and real usability problems. Our proposed method makes extensive use of actual customer-Web site interaction data using Web server logs. We believe that an automated approach to data collection has the potential to resolve some of the problems of current usability evaluation methods: Web server logs can be collected continuously and automatically, enabling on-going Web site evaluations without incurring extraneous costs; and data can be collected for all customers, making it possible to effectively cope with heterogeneity in consumer behavior. 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 should be improved. A method that provides answers to such questions will allow managers to prioritize design and redesign projects to maximize return on investment of the firm's development initiatives.

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

This research closely adheres to the guidelines for conducting design science research (Hevner et al. 2004). Our main design science contribution is the development of a conceptual ***model*** of customer-Web site interaction that enables the assessment of e-commerce Web site effectiveness, a set of measurement ***constructs*** that represent the scope (i.e., *InefficiencyBreadth*) and scale (i.e., *UnitInefficiency*) of usability problems, and an empirical ***method*** that enables the computation of the *InefficiencyBreadth* and *UnitInefficiency* constructs. Collectively these design artifacts provide important contributions and implications to how we think about designing (and redesigning) e-commerce Web sites, in particular, and information systems, more generally.

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 Web sites that support Internet-based selling. The third section outlines the empirical method for Web site evaluation based on this conceptualization. We use *data envelopment analysis* (DEA) as the methodological vehicle for assessing e-commerce Web site usability. We also define two metrics to help identify Web site functionalities that are less than effective in use. *InefficiencyBreadth* quantifies the extent of inefficiency for each different kind of Web site functionality. The *UnitInefficiency* metric computes the relative severity of inefficiency for each of the different Web site 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 Web site. The paper concludes with discussions of the contributions, limitations of the proposed approach, and several areas of extension for future research.

2. A PRODUCTION PERSPECTIVE OF E-COMMERCE WEB SITE EFFECTIVENESS

2.1. A Production Model of Online Shopping

The usability evaluation method proposed in this paper is motivated by our ongoing work on a production economics based model for assessing the effectiveness of e-commerce Web sites for Internet-based selling. We summarize the major elements of the production model of online shopping to provide the conceptual and analytical background for our method for identifying usability problems in e-commerce Web sites.

We conceptualize consumer-Web site interaction during online shopping as a production process in which the customer makes a purchase transaction by utilizing various functionalities provided by the e-commerce Web site.[2] In economics, a *production process* defines the technical means by which inputs (e.g., materials and resources) are converted into outputs (e.g., goods and services) (Varian 1992). 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 Web site. They represent the effort put forth by customers in filling their virtual shopping carts (e.g., the number of product page views, extent of navigation through product listings, references to help pages, etc.). The *outputs* of the production process are the contents of the purchase transaction. For example, the number or dollar value of items

---

[2] This conceptualization is based on literature in service operations management (Chase 1978; Chase and Tansik 1984; Lovelock and Young 1979; Mills and Morris 1986; Walley and Amin 1994), service marketing (Meuter et al. 2000; Zeithaml et al. 1990) and retailing (Ingene 1982; Ingene 1984; Kamakura et al. 1996).

purchased during a shopping trip can be regarded as possible 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 Web site design, and the speed of her Internet connection all could impact how efficient the customer is in producing an online transaction. Borrowing the modeling formalisms of production economics, we represent the online shopping service production process as a *production model* and as its inverse, the *cost model*:[3]

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

$$y_r = f(x_i, s_k, \varepsilon_r^{output})$$

or

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

$$x_i = g(y_r, s_k, \varepsilon_i^{input})$$

where

$f(\cdot)$	=	production function that translates inputs into outputs,
$g(\cdot)$	=	cost function (or inverse production 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 deviations from the production frontier (ε_i^{input}), $i > 0$, $\varepsilon_i^{input}) \geq 0$.

The distinction between the output-oriented production model and the input-oriented cost model is useful due to several reasons. First, the different perspectives provide us with

[3] In management science and operations research, these models are typically referred to as "the primal" and "the dual" – different specifications of the key optimization problem that managers must solve to achieve high value from their technology investments.

flexibility to capture distinctive purchasing behaviors (e.g., goal-directed purchasing vs. experiential 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 involves a consumer who has a target product purchase 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 predetermined level of output, is more appropriate for modeling such goal-directed purchasing behaviors. On the other hand, experiential 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 is a consumer who is shopping for clothes.[4] 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 experiential shopping behaviors.

Second, this distinction is also useful because it provides an analytical basis for interpreting 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. Inefficiencies in the input-oriented cost model (ε_i^{input}) relate to *excess*

[4] We note that the distinction between goal-directed purchasing and experiential shopping is not a pure dichotomy but represents two end points along a continuum of shopping behaviors. For example, shopping for clothes can be quite goal-directed when the exact configuration (i.e., brand, color, size etc.) is predetermined. This has been shown to be the case for some male shoppers (Underhill 1999). So, the example of shopping for clothes as representing experiential shopping behavior is as an illustration.

input: the same amount of outputs could have been produced with less input. Figure 1 shows the basic intuition of the production framework.

INSERT FIGURE 1 ABOUT HERE

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

2.2. Measuring Web Site Effectiveness: A DEA Approach

Conceptualizing online shopping as production enables us to develop a novel perspective for e-commerce Web site evaluation.[5] Since customers are producing a transaction through the e-commerce Web site, the design of the Web site can be viewed as a service production environment. Usability of the Web site thus can be assessed by examining how well the production environment on the e-commerce Web site supports efficient transformation of inputs based on user interaction with the Web site into outputs as purchase transactions. Furthermore, we may utilize frontier estimation methods from production economics – empirical methods typically used for productivity analysis – in evaluating e-commerce Web site performance. Of the various analysis methods available, we selected *data envelopment analysis* (DEA), a linear

[5] 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 the number of errors, 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 conducting primary data collection. Instead, our proposed method focuses on the dimensions of effectiveness and efficiency only.

programming-based non-parametric method for production frontier estimation. We first provide an overview of the analytical method of DEA and then provide details of how DEA can be used to identify usability problems.

2.2.1. Overview of the DEA Method

DEA is a linear programming-based non-parametric method for production frontier estimation. 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).[6]

In DEA, the unit of analysis is called the *decision making unit* or *DMU*. This represents a production unit, which 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. In addition, the analyst may choose between Charnes-Cooper-Rhodes (CCR) and Banker-Charnes-Cooper (BCC) models, depending on whether the production process exhibits constant or

[6] Additional details on DEA and its recent developments can be obtained in recent comprehensive texts such as Charnes et al. (1994), Cooper et al. (2000) and Cooper et al. (2004).

variable returns to scale. The *CCR model* (Charnes et al. 1978; Charnes et al. 1981) structures the analysis with the assumption of 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_{j_0} of DMU j_0, characterized on the basis of the consumption of inputs x_{ij_0} and production of outputs y_{rj_0}, is assessed by solving the linear program shown in Table 1.

INSERT TABLE 1 ABOUT HERE

The first constraint ensures that all observed input combinations lie on or within the production possibility set defined by the production frontier. 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 [7] allows variable returns to scale. The specification of the constraints is such that the production possibilities set conforms to the axioms of production in production economics (i.e., *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

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

$h_j^* = 1$ without any slack output or excess input (i.e., $\varepsilon_r^{output} = \varepsilon_i^{input} = 0, \forall i, r$). All other DMUs with $0 \le h_j^* < 1$ are said to be inefficient (i.e., $\varepsilon_{ij}^{input} > 0$ or $\varepsilon_{rj}^{output} > 0, \exists i, r$). See Figure 2.

INSERT FIGURE 2 ABOUT HERE

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., $h^*_1 = h^*_3 = h^*_4 = 1$). 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_{x_1} and ε_{x_2} less inputs. So DMU_2 exhibits excess inputs of ε_{x_1} for input x_1 and ε_{x_2} for input x_2. The optimal solution when the DEA model is solved for DMU_2, h^*_2, is the ratio of the distance between the origin and DMU'_2 and that between the origin and DMU_2.

2.2.2. Identifying Usability Problems

Efficiency estimation via DEA produces efficiency scores, h^*_j, for each transaction.[8] Hence, we may gain an overall assessment of the effectiveness of the e-commerce Web site 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 Web site is quite effective. However, recall from our earlier discussion that an important managerial concern is to understand not only how the e-commerce Web site is performing, but more importantly, which areas of the Web site are not effective, so as to identify areas for improvement. In such cases, overall efficiency scores do not help us since

[8] 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. We are not particularly interested in the efficiency scores of the DMUs *per se*. Instead, our focus is on the *distribution* of the efficiency scores, which generates insights about the overall *effectiveness* of the production environment in terms of enabling *efficient* production processes.

the efficiency score relates to the productivity of the production environment as a whole (i.e., the e-commerce Web site). Instead we would need to delve deeper into the potential causes by investigating the sources of the observed inefficiencies. One straightforward way to do this is to examine the breadth or scope of observed inefficiencies for each Web site functionality (i.e., how many transactions exhibited inefficiencies with respect to a particular Web site functionality) and the severity or scale of the observed inefficiencies for each Web site functionality (i.e., the level of observed inefficiencies when inefficiencies are observed for a particular Web site functionality). Toward this goal, we define two metrics:

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

The two metrics above can be easily computed from the DEA results. *InefficiencyBreadth* can be calculated by determining the proportion of observations for which input inefficiencies were observed. Since input i in the online shopping production model is conceptualized as the customer's use of Web site functionality i, all non-zero ε_{ij}^{input} represent excess input in the use of Web site functionality i that resulted in the inefficiency in the production of output r. If we define the set $D_i = \left\{ j \in J_i \middle| \varepsilon_{ij}^{input} > 0 \right\}$ (i.e., all DMUs where inefficiency in the use of Web site functionality i was observed), $n_i = |D_i|$ (i.e., the cardinality of D_i, the number of elements or observations in set D_i), and J_i as the total number of DMUs that made use of Web site functionality i, then the proportion of n_i relative to J_i represents the scope of inefficiency due to functionality i: $InefficiencyBreadth_i = n_i / J_i$. *InefficiencyBreadth* is a thus proportional measure. It counts the number of observations or transactions that have inefficiency or excess input, ε_{ij}^{input}, observed with respect to a particular input measure (e.g., Web site functionality i) out of the population of all observations J_i that made use of the Web site functionality i. For instance, if we have a total of 100 observations ($J = 100$), and of those, 80 observations made use of Web site functionality x_1 (e.g., the search function), and of those 80, 20 observations exhibited input inefficiencies for input x_1, usage rate of x_1 (i.e., J_1 / J) would be 80/100 = 0.80 or 80% and $InefficiencyBreadth_1$ would be 20/80 = 0.25 or 25%. This means that 25% of all transactions that actually made use of the search functionality of the Web site showed signs of inefficiencies.

The *UnitInefficiency* metric, representing the severity of observed inefficiencies for a specific Web site functionality, can be determined by analyzing the magnitude of observed input inefficiencies, ε_{ij}^{input}. Since, each observation may have differing levels of outputs, we normalize by output volume: $UnitInefficiency_{irj} = \varepsilon_{ij}^{input} / y_{rj}$. For instance, the first observation ($j = 1$) had input inefficiency with respect to the first Web site 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 this 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 (or standard deviation) etc.) in order to interpret the results. In the next section, we illustrate the utility of our proposed Web site evaluation method by applying it to a real-world operational e-commerce Web site to obtain empirical insights.

3. AN EMPIRICAL APPLICATION

We first discuss the background of our research site and data collection, the specification of our empirical model for the computation of efficiency scores. Then we present the usability insights derived from our analysis.

3.1. Research Site and Data

Data for this study were collected at an online grocery retailer, which requested to remain anonymous. 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 and by mid-July 2000, it had over 9,000 customers who generated more than $16 million in revenue. In 2004, the reported revenue was increased to approximately $65 million. At the time of data collection (mid-2001), the company operated in one metropolitan area, where it was the only e-grocer in its regional market.

Clickstream data were collected directly from the online grocer's Web servers. The Web site uses HTTP session cookies downloaded onto the visitor's computer to track the customer's shopping behavior at the Web site. 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. Web site meta-data were defined to adequately represent the structure and content of the e-commerce Web site.[9] This helped to make sense of the Web site usage behaviors by identifying the nature and content of the pages the customers have gone through to browse and purchase products on the Web site. Web site usage metrics were then extracted to measure the extent to which various areas of the Web site 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. The data integration process is presented in Figure 3.

INSERT FIGURE 3 ABOUT HERE

3.2. DEA Model Specification

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 Web site. 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.

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 Web site is the time-pinched customer who seeks convenience in her grocery shopping activities. Consequently, one of the key operational goals for Web site design set forth by the managers at

[9] Web site structural meta-data provide the topology of the Web site (e.g., entry points, the product hierarchy, link structures, etc.) that can be represented as a directed graph structure. Web site meta-data also included semantic information related to the content on the individual pages, which we represented as keyword matrices.

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 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; Charnes et al. 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 production scale, is under the control of the customer and not the e-commerce firm. If an optimal production scale exists, one cannot enforce it.[10] 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 exhibits variable returns to scale, hence the BCC model formulation is appropriate.[11]

Once the model orientation and specification have been determined, the next step is to select the appropriate input and output variables for empirical estimation. 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. In online shopping, *inputs* consist of

[10] Enforcing a production scale on a consumer is analogous to suggesting that she purchase more (or less) items because she would be more efficient with a different transaction size.
[11] Details of the returns-to-scale analyses are not shown here, but are available upon request.

customers' use of various Web site functionalities and the *output* consists of a checkout of a basket of products.[12] The input and output variables are summarized in Table 2.

INSERT TABLE 2 ABOUT HERE

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

3.3. Empirical Results

3.3.1. Results of Overall DEA Efficiency Scores

Figure 4 shows the efficiency scores of all DMUs (J = 5,383) against the respective output of each observation. The horizontal axis represents the efficiency scores of the online shopping transactions ($0 \leq h_j^* \leq 1$). The output level (i.e., number of items in 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 Web site efficiency, with varying efficiency scores at all output levels, suggesting that the Web site may not be entirely effective and there is room to improve.

INSERT FIGURE 4 ABOUT HERE

3.3.2. Inefficiency Results by Web Site Functionality

To gain insights into the potential causes of the observed overall Web site inefficiency, we analyzed the inefficiencies by Web site functionality with the inefficiency metrics proposed earlier. (See Table 3 for actual measures and Figure 5 for a graphical summary.)

[12] Following recommendations in the DEA literature (e.g., Thomas et al. 1998), the conceptualization and operationalization of input and output variables was determined via in-depth discussion with the managers at the online grocer.

INSERT TABLE 3 AND FIGURE 5 ABOUT HERE

We first observe that not all Web site functionalities are used to an equal extent. (See the 4th column in Table 3 and gray bars in Figure 5.) Some are almost always used in purchase transactions (e.g., *Checkout* – 100%[13] and *ProductList* – 99.94%). However, we also see that several functionalities are used at a very low rate, with less than one-fifth of the transactions making use of such functionalities (e.g., *Help* – 11.70%, *OrderHistory* – 17.22% and *Recipe* – 18.69%), while others are fairly extensively used (but not all the time), with over two-thirds of the transactions making use of such functionalities (e.g., *Search* – 68.31%, *Promotion* – 68.98%, *PersonalList* – 74.77% and *ProductInformation* – 81.83%). Recall that $InefficiencyBreadth_i$ (5th column in Table 3 and white bars in Figure 5) measures the proportion of DMUs for which excess input for Web site functionality i was observed. For example, the first Web site functionality (*ProductInformation*, DEA input variable x_1) was used by 4405 DMUs (i.e., purchase transactions) of which excess input was observed in 2,272 cases ($n_1 = 2272$). Hence, $InefficiencyBreadth_1$ for *ProductInformation* is 51.58% (i.e., 2272 / 4405 = 0.5158). We see that of the various Web site functionalities, *InefficiencyBreadth* was greatest for *Help* (98.6%), then *Promotion* (57.1%), *ProductInformation* (51.6%), *Recipe* (49.6%), *PersonalList* (42.0%) and *Search* (37.5%). With the remaining three Web site functionalities (i.e., *OrderHistory*, *ProductList*, and *Checkout*), the breadth of inefficiency was less significant.

Even though *InefficiencyBreadth* provides a useful metric that represents how widespread inefficiencies due to specific Web site functionality were, it does not provide much information as to the severity of those observed inefficiencies. We also investigate the *UnitInefficiency*

[13] Since we are capturing completed transactions, the checkout functionality (*Checkout*) is inevitably always used.

metric to gain more insights. (See right-most column in Table 3 and the line data in Figure 5.) The results show that the severity of inefficiencies were greatest for Web site functionalities *ProductList* (0.385), *Search* (0.342) and *ProductInformation* (0.337), moderate for *PersonalList* (0.261), *Promotion* (0.176), *Recipe* (0.158), *Checkout* (0.143) and *OrderHistory* (0.125), and low for *Help* (0.035).

Interpreting the metrics together provide a more holistic picture of Web site usage and effectiveness. The combined results suggest that Web site functionalities *ProductInformation*, *PersonalList*, *Promotion* and *Search* were most problematic in incurring inefficiencies at the e-tailer's Web site. These were the functionalities that exhibited extensive use (i.e., over 66% in usage rate), widespread inefficiencies (i.e., *InefficiencyBreadth* measures ranging between approximately 40-60%) and also severe inefficiencies (i.e., average *UnitInefficiency* measures ranging between 0.17 and 0.34). On the other hand, although, *Checkout* and *ProductList* were used in almost all transactions (i.e., usage rate ≈ 100%), only a very small proportion of transactions exhibited inefficiencies in these functionalities even though the severity of the inefficiencies (when observed) was moderate to high. In contrast, inefficiencies with *Help* were widespread, but this was rarely used and the severity of the observed inefficiency was negligible.

We may also formally test the differences in the *UnitInefficiency* scores between the Web site 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. The statistical procedure involves testing whether the means of the inefficiency score probability distributions for different conditions are different based on two test statistics. The application of each depends on whether inefficiency deviations of the observed data are assumed to be drawn from an *exponential* or a *half-normal distribution.*[14]

The 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 G_i is denoted by $\theta_j^{G_i}$. 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 Web site functionalities, say a and b) is that inefficiencies due to the two Web site 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 Web site functionality a is greater than those due to Web site functionality b (i.e., Web site functionality a is showing more inefficiencies than Web site functionality b). The test statistic is:

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

. 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 G_a and G_b and $n = n_a + n_b$. On the other hand, if one assumes the inefficiency deviations to be half-normally distributed then a different test statistic is used: $\frac{\sum_{j \in D_a} (\theta_j^{G_a} - 1)^2 \big/ n_a}{\sum_{j \in D_b} (\theta_j^{G_b} - 1)^2 \big/ n_b}$. This statistic again asymptotically follows an F-distribution with (n_a, n_b) degrees of freedom for large values of n. Using this statistical testing method, we conducted pair-wise *UnitInefficiency* score comparisons for each of the Web site functionalities. The results are summarized in Table 4.

INSERT TABLE 4 ABOUT HERE

[14] 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.

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 Web site 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 Web site 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 Web site 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.[15]

Until now, we have presented general results from using the Web site evaluation method to identify potentially problematic Web site 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 Web site 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 Web site functionalities are problematic for different groups.

---

[15] Note that we are naively presenting these statistical test results without making use of additional background information that managers would typically have in practice. For example, although we find that the Web site functionalities of *Product Information*, *Search* and *Personal List* are the most problematic, the managers of the Web site may not find this result particularly alarming if they have reason to believe that those functionalities are inherently inefficient. In other words, Web site managers 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 Web site functionalities.

3.3.3. Inefficiency Results by Output Volume

To further demonstrate the value of the proposed evaluation method, we present additional results from a simple extension: investigating Web site 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 Web site interaction behaviors from those that conduct lower volume ones, we analyzed the DEA results to explore these issues.

Figure 6 shows the distributions of *UnitInefficiency* values by output volume for each of the Web site functionalities. Several results are noteworthy.

INSERT FIGURE 6 ABOUT HERE

The distribution of *UnitInefficiency* values is skewed with most observations in the lower range (with a long tail). As expected, *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 Web site functionalities *ProductInformation*, *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 Web site functionality. This suggests that the Web site may be geared toward supporting smaller carts more effectively than larger ones.

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 Web site is quite typical in Internet-based selling, with hierarchical product categories for drill-down navigation, 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 per transaction is small (i.e., cart size of 1 to 5 line-items), a different overall design strategy may be required for grocery shopping where the number of different items being purchased is substantially larger (i.e., cart size of 40+ line-items).

4. CONCLUSION

The success of e-commerce retailers hinges largely on their ability to provide a high-quality Web site. So e-commerce retailers need to constantly monitor the effectiveness of their Internet-based storefronts to identify those Web site areas that are problematic. However, evaluating the effectiveness of e-commerce Web site design is a complex problem for e-commerce retailers. Furthermore, current evaluative methods do not offer practical and effective means for a solution to this problem. We proposed an innovative method for automated Web site usability assessment and identification of problem areas.

4.1. Contributions

By modeling online shopping as an economic production process and using evaluative methods for frontier analysis, we developed a productivity-based model for Web site effectiveness and a set of metrics that help us to identify problematic areas within the site. We also demonstrated the value of our method by applying it to the evaluation of a real-world e-commerce Web site. Through the application, it was possible to gain a deeper understanding of which Web site 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 Web site. Such knowledge discovery provides a useful starting point for delving deeper into these design issues.

The proposed Web site evaluation method provides significant benefits over other methods that are currently widely used. Our empirical insights could not have been uncovered using the traditional empirical 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 Web site. In fact, the problem of scalability is a major concern with the extant evaluation methods. With user testing, for instance, deciding how many subjects will be needed to produce a representative picture of Web site 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 Web site 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). A practical application scenario would be to use our proposed method to identify which areas of the Web site are most problematic and employ other methods such as testing or inquiry on these areas to understand why usability problems are arising.

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 frequent Web site redesigns, it becomes difficult to solicit continuous responses for each redesign cycle. 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. On the contrary, 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 Web site. 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 usability metrics so that e-commerce managers can continuously monitor the effectiveness of their Web site without incurring the costs of extraneous data collection and tedious analysis.

4.2. Limitations

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 Web site evaluation method can be inferred by the interesting analyses and 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 Web site functionalities (e.g., *ProductInformation*, *PersonalList*, *Promotion* and *Search*) were more problematic than others (e.g., *Help, ProductList* and *Checkout*). However, the reader should not over-generalize and assume that these Web site functionalities would also be problematic on other e-commerce Web sites. 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 Web site). Instead, the focus of the proposed evaluation method is to provide to managers at e-commerce firms useful feedback concerning how their customers are *actually* performing in the presence of their *current* Web site design.

Furthermore, it is important to note that the data collection window of two weeks may have introduced bias in the dataset; only transactions completed within the two-week window are included. This is an important concern since many online consumers engage in "look-to-book" type shopping.[16] Consequently, the results of the current analyses need to be interpreted with this limitation in mind. That said, our method could be applied with a more complex model that tracks consumer purchases for a longer period so that such situations might also be handled.

Another limitation with respect to the interpretation of the results 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 Web site functionalities. Our future research agenda includes extending the online shopping model so that such interaction effects can also be dealt with.

We also acknowledge potential concerns relating to the applicability of our proposed method to other contexts of Internet-based selling. In our currently study, we have applied our method to gain insights into the effectiveness of the Web site of an online grocer. An important 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

[16] *Look-to-book 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.

input-oriented production framework, which is more appropriate for goal-directed purchasing. Other e-commerce Web sites 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 Web sites. For instance, in the case of Internet-based selling of clothes, consumer purchase behaviors will typically entail experiential 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, Web site navigation behaviors that results in more products identified, considered and purchased (i.e., output) given the same amount of Web site use (i.e., input) would be regarded as more effective. The proposed model and method is general enough so these different types of consumer behaviors can be appropriately captured.

Even if the proposed usability evaluation method can be applied to Internet-based selling Web sites that deal with different types of products and shopping motives, we also need to think about whether the proposed method can effectively be applied to non-commercial Web site applications (e.g., search engines, online portals, informational Web sites, etc.) We believe that our proposed method works so long as the perspective analogy is appropriate and applicable. So, if it is appropriate to model HCI behavior as a production process (i.e., as an input-output mapping, even if the input are not "consumed" in the usual sense in production), our method will apply. In fact, 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 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 used here. We acknowledge, however, that the production framework applies more easily in the context of transactional e-commerce Web sites due to the relatively unambiguous characterization of inputs, and more importantly outputs (e.g., products purchased). For informational Web sites, 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 consumer purchases an item, or adds the item to her virtual shopping cart, then we can reasonably assume that the person was seeking that item. Innovative techniques for inferring user needs from observable clickstream behaviors (e.g., Chi et al. 2001) should prove valuable in these regards.

Finally, the proposed evaluation method is intended to be used as a tool for the continuous management of Web site 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, our valuation method is intended for use within a firm for managing its Web site development initiatives.

Given such use contexts, we need to discuss the broader applicability issues within the context of systems development for Internet-based selling. Recall that one of the major goals of this research is to propose a Web site evaluation technique that provides *automated analysis* capabilities to help deal with faster software development life cycles by minimizing (or at least reducing) the evaluation costs at each redesign. But *automation* does not come for free. For such automation to seamlessly occur, the Web site development process must incur additional costs of designing the user interface and application logic so that the log data that is automatically collected is in a structure and form that can be readily analyzed. For example, developers may need to incorporate additional flags in the Web application code so that it is easier to automatically identify customers, transactions, sessions, page views and also content (i.e., products) within the Web pages shown. Such additional activities may increase the level of complexity in the design and consequently increase the difficulty and costs of functional testing and debugging. There are even firms that believe that the world is changing so fast that any attempt at analysis is fruitless since it would deter them from quickly adapting to take advantage of emerging market opportunities. However, as Andreas Weigend, former chief scientist at Amazon.com, notes, the benefits of making better decisions based on careful analyses of implicit behavioral data (e.g., clickstream data) far outweigh the extra costs of developing and maintaining information systems that support such analyses (Wegner 2005). We hope that our proposed method will help firms make better decisions.

4.3. Coda: Evaluating This Research from a Design Science Perspective

The scientific value of any research needs to be rigorously assessed. Furthermore, research following different epistemological philosophies (e.g., positivism vs. interpretivism) or paradigms (e.g., behavioral science vs. design science) should be evaluated according to standards that match the philosophy or paradigm of the research conducted. In their recent article, Hevner et al. (2004) present a conceptual framework and guidelines for understanding, executing and evaluating design science research in Information Systems. In conclusion, we candidly evaluate our current research in light of the seven guidelines presented in Hevner et al. (2004), and argue that this paper exemplifies rigorous design science research.

- **Guideline 1 (Design as an Artifact).** "*Design science research must produce a viable artifact in the form of a construct, a model, a method or an instantiation.*"

The current research has successfully produced several viable IT artifacts. We develop two new metrics (i.e., *InefficiencyBreadth* and *UnitInefficiency*), ***constructs*** that provide a concise representation and vocabulary for understanding and interpreting overall Web site usage and usability problems. These metrics represent the scope (*InefficiencyBreadth*) and scale (*UnitInefficiency*) of usability problems that can be associated with a particular Web site functionality. These constructs are based upon our ***model*** of online shopping as economic production which provides the theoretical foundation for assessing Web site effectiveness in terms of productivity, efficiency and/or inefficiency. Furthermore, we propose an empirical ***method*** for computing the proposed metrics. While we do not develop a fully functional prototype that ***instantiates*** the proposed constructs, model and methods, we demonstrate the utility of our design artifacts by applying them to a real-world e-tailer.

- **Guideline 2 (Problem Relevance).** *"The objective of design science research is to develop technology-based solutions to important and relevant business problems."*

Designing usable Internet-based selling Web sites is extremely important and relevant to e-commerce firms. We have argued that current methods for assessing the usability of e-commerce Web sites have some major shortcomings (e.g., excessive cost requirements, lack of scalability, potential for bias in results etc.) and are limited when it comes to rapidly providing managerially-actionable advice to managers. We identified viable opportunities for solving these urgent managerial problems with automated tools for usability evaluation. Our proposed method fills an important niche in the arsenal of available tools for e-commerce managers.

- **Guideline 3 (Design Evaluation).** *"The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods."*

Our major design artifact is essentially a method for usability evaluation. So the most relevant evaluation is to compare its performance against other methods for usability evaluation. However, direct comparison is difficult given that the purpose of our proposed method is different from existing approaches. The purpose of our proposed method is to identify potential areas of usability problems from large-scale data so that managers can direct their attention to particular problem issues and areas. Traditional methods (e.g., user testing, inquiry and/or inspection) focus more on the qualitative nature of usability problems. In other words, if the emphasis of traditional methods is on understanding *why* some usability problems occur, then the focus of our method is to quickly identify *where* the usability problems might be occurring.

Accordingly, we follow observational and descriptive approaches to evaluation. We applied our method to the evaluation of the usability of a real-world online retailer and were able to uncover interesting insights that would have been difficult with traditional approaches. We also made informed arguments about how our proposed method overcomes some of the shortcomings of prior approaches (e.g., automated data collection and computation of metrics reduce evaluation costs, unobtrusive data collection reduces potential for bias in evaluation results, use of large-scale data about all customers solves the scalability problems, etc.)

- **Guideline 4 (Research Contributions).** "*Effective design science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies."*

The major contribution of our research is a novel way to think about the effectiveness of Internet-based selling Web sites, in particular, and information systems, more generally. The adoption of the *production paradigm* for understanding systems effectiveness opens up interesting directions for future research. For example, the long-standing literature in production economics has many concepts (e.g., marginal rate of substitution, technical vs. scale inefficiencies, etc.), analytical tools and methods that we have yet to apply in this context. Future research will benefit from our exploration.

- **Guideline 5 (Research Rigor).** *"Design science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact."*

Our research combines rigorous elements from multiple disciplines such as economics (e.g., production economics, production model, cost model), management science (e.g., data envelopment analysis), marketing (e.g., online consumer behavior, purchase decision making), HCI (e.g., usability engineering), computer science (e.g., data mining and Web usage mining) and IS (e.g., systems development life cycle, iterative development). We draw from these rigorous knowledge bases in the development of our proposed usability evaluation method.

- **Guideline 6 (Design as a Search Process).** *"The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment."*

Our research also follows an effective search process. The conceptualization and formulation of the design objective utilized means-ends analysis by first clearly defining the goal state (i.e., the ability to inform managers whether their online storefront is effective, and if not, which areas are problematic and require attention). Our research also identified opportunities (e.g., the availability and potential value of large-scale usage traces from Web server logs) as well as the constraints (i.e., the common denominator of available data such as Web server logs, transaction data and customer data) to define the design requirements (i.e., a method for automatically identifying usability problems in e-commerce Web sites).

- **Guideline 7 (Communication of Research).** *"Design science research must be presented effectively both to technology-oriented as well as management-oriented audiences"*

We intend that the current research is presented for both technology-oriented as well as managerially-oriented audiences. For the technology-oriented, we include the formal specification of our model (i.e., the production model) and empirical computation (e.g., the DEA formulation for efficiency computation) of the metrics. We have shown that the proposed method can be readily implemented. For the managerially-oriented, we have strived to clearly outline what additional investments (e.g., additional costs of programming the measurement logic into application servers) are required to adopt and implement our proposed method. In addition, we also clearly highlight the value (and the limits of value) of our proposed method by showing what kinds of analyses can (or cannot) be performed and what kinds of insights may (or may not) be generated from such analyses. Managers should be able to make informed decisions about whether the adoption, implementation and use of our method will generate value.

REFERENCES

Agarwal, R., and Venkatesh, V. "Assessing a Firm's Web Presence: A Heuristic Evaluation Procedure for the Measurement of Usability," *Information Systems Research* (13:2), 2002, pp. 168-186.

Aigner, D.J., and Chu, S.F. "On Estimating the Industry Production Function," *American Economic Review* (58:4), 1968, pp. 826-839.

Albert, T.C., Goes, P.B., and Gupta, A. "GIST: A Model for Design and Management of Content and Interactivity of Customer-Centric Web Sites," *MIS Quarterly* (28:2), 2004, pp. 161-182.

Anderson, J., Fleak, F., Garrity, K., and Drake, F. "Integrating Usability Techniques into Software Development," *IEEE Software* (18:1), 2001, pp. 46-53.

Andrews, D., Nonnecke, B., and Preece, J. "Electronic Survey Methodology: A Case Study in Reaching Hard-to-Involve Internet Users," *International Journal of Human-Computer Interaction* (16:2), 2003, pp. 185-210.

Babin, B.J., Darden, W.R., and Griffin, M. "Work and/or Fun: Measuring Hedonic and Utilitarian Shopping Value," *Journal of Consumer Research* (20:4), 1994, pp. 644-656.

Bakos, J.Y. "The Emerging Landscape for Retail E-Commerce," *Journal of Economic Perspectives* (15:1), 2001, pp. 69-80.

Banker, R.D. "Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation," *Management Science* (39:10), 1993, pp. 1265-1273.

Banker, R.D., Charnes, A., and Cooper, W.W. "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis," *Management Science* (30:9), 1984, pp. 1078-1092.

Banker, R.D., Datar, S.M., and Kemerer, C.F. "A Model to Evaluate Variables Impacting the Productivity of Software Maintenance Projects," *Management Science* (37:1), 1991, pp. 1-18.

Banker, R.D., and Kauffman, R.J. "Reuse and Productivity: An Empirical Study of Integrated Computer-Aided Software Engineering (ICASE) at the First Boston Corporation," *MIS Quarterly* (15:3), 1991, pp. 374-401.

Banker, R.D., and Slaughter, S.A. "A Field Study of Scale Economies in Software Maintenance," *Management Science* (43:12), 1997, pp. 1709-1725.

Bevan, N., Barnum, C., Cockton, G., Nielsen, J., Spool, J.M., and Wixon, D. "Panel: The 'Magic Number 5:' Is It Enough for Web Testing?," in *Proceedings of the 2003 ACM Conference on Human Factors in Computing Systems*, G. Cockton and P. Korhonen (eds.), ACM Press, New York: NY, Ft. Lauderdale, FL, April 5-10, 2003, pp. 698-699.

Bias, R.G. "Usability Triage for Web Sites," in *Proceedings of the 6th Conference on Human Factors and the Web*, P. Kortum and E. Kunzinger (eds.), Austin, TX, June 19, 2000.

Bloch, P.H., Sherrell, D.L., and Ridgway, N.M. "Consumer Search: An Extended Framework," *Journal of Consumer Research* (13:1), 1986, pp. 119-126.

Brajnik, G. "Automatic Web Usability Evaluation: What Needs to Be Done?," in *Proceedings of the 6th Conference on Human Factors and the Web*, P. Kortum and E. Kunzinger (eds.), Austin, TX, June 19, 2000.

Byrne, M.D., Wood, S.D., Sukaviriya, P.N., Foley, J.D., and Kieras, D.E. "Automating Interface Evaluation," in *Proceedings of the 1994 ACM Conference on Human Factors in Computing Systems*, B. Adelson, S. Dumais and J. Olson (eds.), ACM Press, New York: NY, Boston, MA, April 24-28, 1994, pp. 232-237.

Card, S.K., Moran, T.P., and Newell, A. *The Psychology of Human Computer Interaction*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1983.

Charnes, A., Cooper, W.W., Lewin, A.Y., and Seiford, L.M. *Data Envelopment Analysis: Theory, Methodology, and Applications*, Kluwer Academic Publishers, Boston, MA, 1994.

Charnes, A., Cooper, W.W., and Rhodes, E. "Measuring Efficiency of Decision-Making Units," *European Journal of Operational Research* (2:6), 1978, pp. 428-449.

Charnes, A., Cooper, W.W., and Rhodes, E. "Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through," *Management Science* (27:6), 1981, pp. 668-697.

Chase, R.B. "Where Does the Customer Fit in a Service Operation?" *Harvard Business Review* (56:6), 1978, pp. 138-139.

Chase, R.B., and Tansik, D.A. "The Customer Contact Model for Organization Design," *Management Science* (29:9), 1984, pp. 1037-1050.

Chen, P.-Y., and Hitt, L.M. “Measuring Switching Costs and the Determinants of Customer Retention in Internet-Enabled Businesses: A Study of the Online Brokerage Industry,” *Information Systems Research* (13:3), 2002, pp. 255-274.

Chi, E.H., Pirolli, P.L.T., Chen, K., and Pitkow, J.E. “Using Information Scent to Model User Information Needs and Actions on the Web,” in *Proceedings of the 2001 ACM Conference on Human Factors in Computing Systems*, J.A. Jacko, A. Sears, M. Beaudouin-Lafon and R.J.K. Jacob (eds.), ACM Press, New York: NY, Seattle, WA, March 31-April 5,2001, pp. 490-497.

Chidamber, S.R., Darcy, D.P., and Kemerer, C.F. “Managerial Use of Metrics for Object Oriented Software: An Exploratory Analysis,” *IEEE Transactions on Software Engineering* (24:8), 1998, pp. 629-639.

Chin, J.P., Diehl, V.A., and Norman, K.L. “Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface,” in *Proceedings of the 1988 ACM Conference on Human Factors in Computing Systems*, J.J. O’Hare (ed.), ACM Press, New York: NY, Washington, D.C., May 15-19, 1988, pp. 213-218.

Cooley, R., Mobasher, B., and Srivastava, J. “Data Preparation for Mining World Wide Web Browsing Patterns,” *Journal of Knowledge and Information Systems* (1:1), 1999, pp. 5-32.

Cooper, W.W., Seiford, L.M., and Tone, K. *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References, and DEA-Solver Software*, Kluwer Academic Publishers, Boston, MA, 2000.

Cooper, W.W., Seiford, L.M., and Zhu, J. *Handbook on Data Envelopment Analysis*, Kluwer Academic Publishers, Boston, MA, 2004.

Cutler, M., and Sterne, J. “E-Metrics: Business Metrics for the New Economy,” NetGenesis, Cambridge, MA.

Dray, S.M. “The Importance of Designing Usable Systems,” *Interactions* (2:1), 1995, pp. 17-20.

Drott, M.C. “Using Web Server Logs to Improve Site Design,” in *Proceedings of the 16th ACM Conference on Systems Documentatio*n, Quebec, Canada, September 24-26, 1998, pp. 43-50.

Duggan, E.W. “Generating Systems Requirements with Facilitated Group Techniques,” *Human-Computer Interaction* (18:4), 2003, pp. 373-394.

Dumas, J.S., and Redish, J.C. *A Practical Guide to Usability Testing*, (revised ed.), Intellect Books, Exeter, UK, 1999.

Fu, L., Salvendy, G., and Turley, L. “Effectiveness of User Testing and Heuristic Evaluation as a Function of Performance Classification,” *Behaviour and Information Technology* (21:2), 2002, pp. 137-143.

Fuller, T., and de Graaff, J.J. “Measuring User Motivation from Server Log Files,” in *Proceedings of the 2nd Conference on Human Factors and the Web*, M. Czerwinski (ed.), Redmond, WA, October 30,1996.

Gallaugher, J. “Challenging the New Conventional Wisdom of Net Commerce Strategies,” *Communications of the ACM* (42:7), 1999, pp. 27-29.

Hevner, A.R., March, S.T., Park, J., and Ram, S. “Design Science Research in Information Systems,” *MIS Quarterly* (28:1), 2004, pp. 75-105.

Ingene, C.A. “Labor Productivity in Retailing,” *Journal of Marketing* (46:4), 1982, pp. 75-90.

Ingene, C.A. “Productivity and Functional Shifting in Spatial Retailing: Private and Social Perspectives,” *Journal of Retailing* (60:3), 1984, pp. 15-36.

International Organization for Standardization (ISO) “Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 11: Guidance on Usability, ISO 9241-11,” Geneva, Switzerland, 1998.

Ivory, M.Y., and Hearst, M.A. “The State of the Art in Automating Usability Evaluation of User Interfaces,” *ACM Computing Surveys* (33:4), 2001, pp. 470-516.

John, B.E., and Kieras, D.E. “The GOMS Family of User Interface Analysis Techniques: Comparison and Contrast,” *ACM Transactions on Computer-Human Interaction* (3:4), 1996, pp. 320-351.

Johnson, E.J., Bellman, S., and Lohse, G.L. “Cognitive Lock-in and the Power Law of Practice,” *Journal of Marketing* (67:2), 2003, pp. 62-75.

Kamakura, W.A., Lenartowicz, T., and Ratchford, B.T. “Productivity Assessment of Multiple Retail Outlets,” *Journal of Retailing* (72:4), 1996, pp. 333-356.

Keeney, R.L. “The Value of Internet Commerce to the Customer,” *Management Science* (45:4), 1999, pp. 533-542.

Lovelock, C.H., and Young, R.F. “Look to Consumers to Increase Productivity,” *Harvard Business Review* (57:3), 1979, pp. 168-178.

Meuter, M.L., Ostrom, A.L., Roundtree, R.I., and Bitner, M.J. “Self-Service Technologies: Understanding Customer Satisfaction with Technology Based Service Encounters,” *Journal of Marketing* (64:3), 2000, pp. 50-64.

Mills, P.K., and Morris, J.H. “Clients As ‘Partial’ Employees of Service Organizations: Role Development in Client Participation,” *Academy of Management Review* (11:4), 1986, pp. 726-735.

Moe, W.W., and Fader, P.S. “Uncovering Patterns in Cybershopping,” *California Management Review* (43:4), 2001, pp. 106-117.

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

Nielsen, J., and Molich, R. “Heuristic Evaluation of User Interfaces,” in *Proceedings of the 1990 ACM Conference on Human Factors in Computing Systems*, J.C. Chew and J. Whiteside (eds.), ACM Press, New York: NY, Seattle, WA, April 1-5, 1990, pp. 249-256.

Pirolli, P.L.T. “Computational Models of Information Scent-Following in a Very Large Browsable Text Collection,” in *Proceedings of the 1997 ACM Conference on Human Factors in Computing Systems*, S. Pemberton (ed.), ACM Press, New York: NY, Atlanta, GA, March 22-27, 1997, pp. 3-10.

Pirolli, P.L.T., and Card, S.K. “Information Foraging,” *Psychological Review* (106:4), 1999, pp. 643-675.

Rosenstein, M. “What Is Actually Taking Place on Web Sites: E-Commerce Lessons from Web Server Logs,” in *Proceedings of the 2nd ACM Conference on Electronic Commerce,* Minneapolis, MN, October 17-20, 2000, pp. 38-43.

Schubert, P., and Selz, D. “Web Assessment: Measuring the Effectiveness of Electronic Commerce Sites Going Beyond Traditional Marketing Paradigms,” in *Proceedings of the 32nd Hawaii International Conference on System Sciences*, R.H. Sprague (ed.), IEEE Computer Society Press, Los Alamitos: CA, Maui, HI, January 5-8, 1999.

Schwarz, N. “Self-Reports: How the Questions Shape the Answers,” *American Psychologist* (54:2), 1999, pp. 93-105.

Spool, J.M., Scanlon, T., Schroeder, W., Synder, C., and DeAngelo, T. *Web Site Usability: A Designer’s Guide*, Morgan Kaufmann Publishers, San Francisco, CA, 1999.

Spool, J.M., and Schroeder, W. "Testing Web Sites: Five Users Is Nowhere near Enough," in *Proceedings of the 2001 ACM Conference on Human Factors in Computing Systems*, J.A. Jacko and A. Sears (eds.), ACM Press, New York: NY, Seattle, WA, Mar 31-April 5, 2001, pp. 285-286.

Thomas, R.R., Barr, R.S., Cron, W.L., & Slocum, J.W., Jr. "A Process for Evaluating Retail Store Efficiency: A Restricted DEA Approach." *International Journal of Research in Marketing* (15:5), 1998, pp. 487-503.

Underhill, P. *Why We Buy: The Science of Shopping,* Touchstone, New York: NY, 1999.

Varian, H.R. *Microeconomic Analysis*, (3rd ed.), Norton, New York: NY, 1992.

Walley, P., and Amin, V. "Automation in a Customer Contact Environment," *International Journal of Operations and Production Management* (14:5), 1994, pp. 86-100.

Wegner, J. "I Search, Therefore I Am," in: SASCOM, 2005, pp. 37-41.

Wharton, C., Rieman, J., Lewis, C., and Polson, P.G. "The Cognitive Walkthrough Method: A Practitioner's Guide," in *Usability Inspection Methods*, J. Nielsen and R.L. Mack (eds.), Wiley, New York: NY, 1994, pp. 105-140.

Zeithaml, V.A., Parasuraman, A., and Berry, L.L. *Delivering Quality Service: Balancing Customer Perceptions and Expectations*, Free Press, New York: NY, 1990.

TABLES AND FIGURES

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 $\forall j$ observations	$\lambda_j \geq 0$, for $\forall j$ observations
$\sum_{j=1}^{J} \lambda_j = 1$, $j = 1,\ldots,J$ observations	$\sum_{j=1}^{J} \lambda_j = 1$, $j = 1,\ldots,J$ observations

Table 2. DEA Model's Input and Output Variables

Category	Variable	Measure	Description (Specified as # of …)
Inputs	x_1	*Products*	Product page views
	x_2	*Lists*	Product list views
	x_3	*Personal*	Personal list views
	x_4	*OrderHistory*	Order history page views
	x_5	*Search*	Search conducted
	x_6	*Promotion*	Promotional page views
	x_7	*Recipe*	Recipe page views
	x_8	*Checkout*	Checkout pages
	x_9	*Help*	Help page views
Output	y_1	*BasketSize*	Different products at checkout

Table 3. Inefficiency of Web Site Functionality

Web Site Functionality	J_i	n_i	*Usage* (J_i / J)	*Inefficiency Breadth*	*UnitInefficiency (Average)*
ProductInformation (x_1)	4405	2272	81.83%	51.58%	0.3368
ProductList (x_2)	5380	359	99.94%	6.67%	0.3851
PersonalList (x_3)	4025	1690	74.77%	41.99%	0.2609
OrderHistory (x_4)	927	128	17.22%	13.81%	0.1248
Search (x_5)	3677	1378	68.31%	37.48%	0.3415
Promotion (x_6)	3713	2121	68.98%	57.12%	0.1763
Recipe (x_7)	1006	499	18.69%	49.60%	0.1580
Checkout (x_8)	5383	240	100.00%	4.46%	0.1427
Help (x_9)	630	621	11.70%	98.57%	0.0348

Table 4. Statistical Pair-Wise Comparison of UnitInefficiency for Web Site Functionalities

Site Functionalities	**Site 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 Web site functionality of the row and the Web site 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*.

Figure 1. Efficiency Frontiers

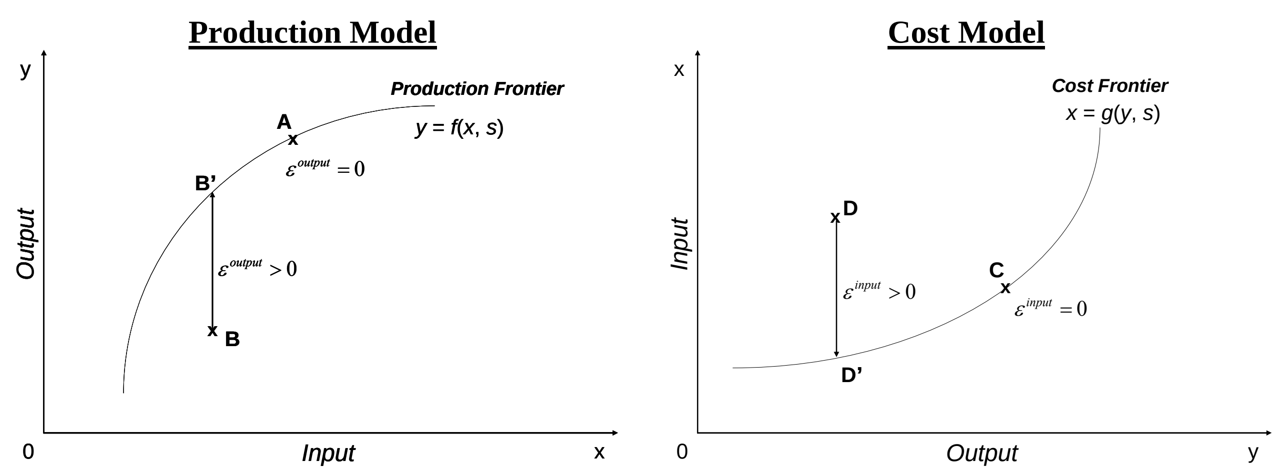
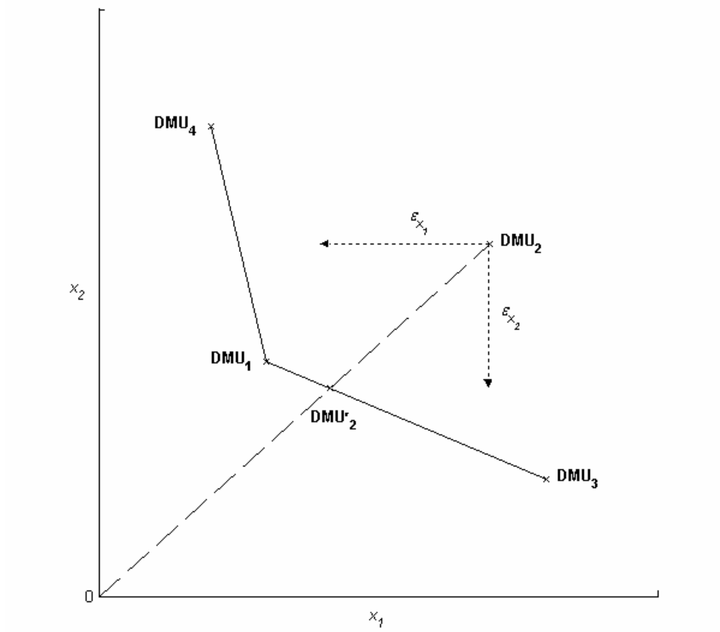


Figure 2. DEA Production Frontier and Production Inefficiency (two inputs, one output)



Note: The graph represents an *output isoquant*. Hence, all data are normalized for unit output reflecting the product of a similar amount of output. The inputs, x_1 and x_2, are shown as the axes.

Figure 3. Data Integration

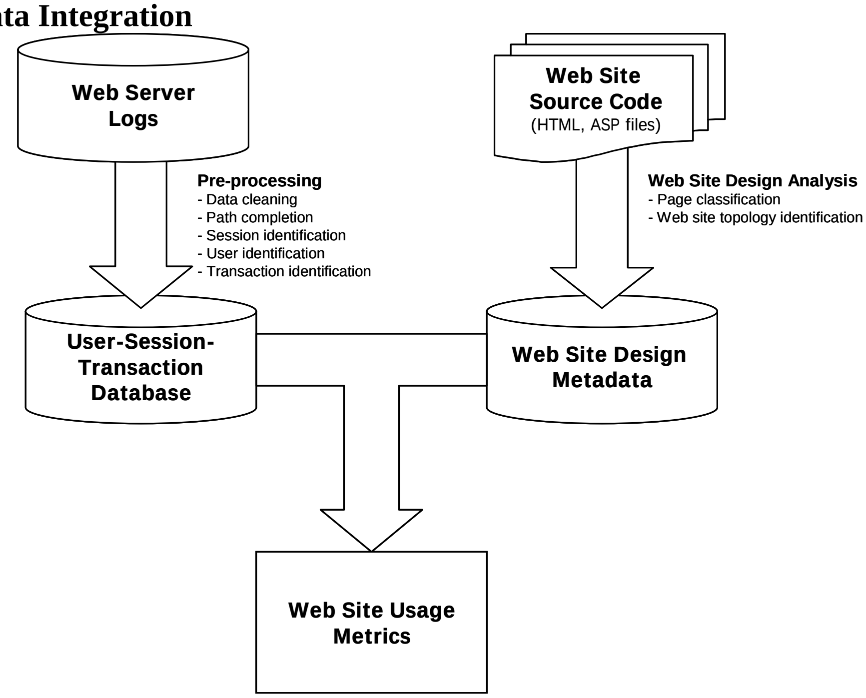


Figure 4. DEA Efficiency Scores by Output Level for Shopping Basket Size at Checkout

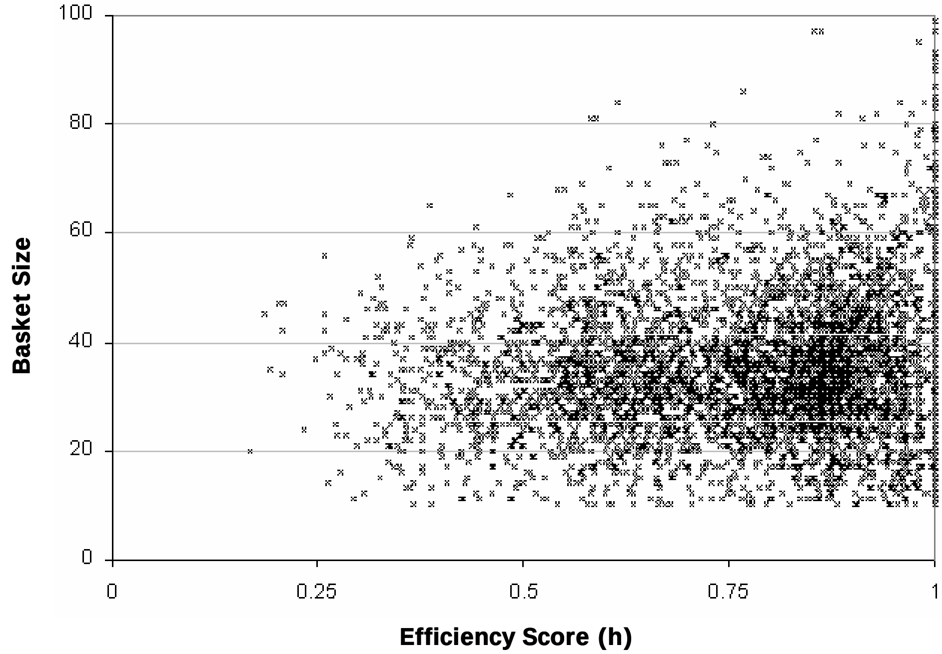
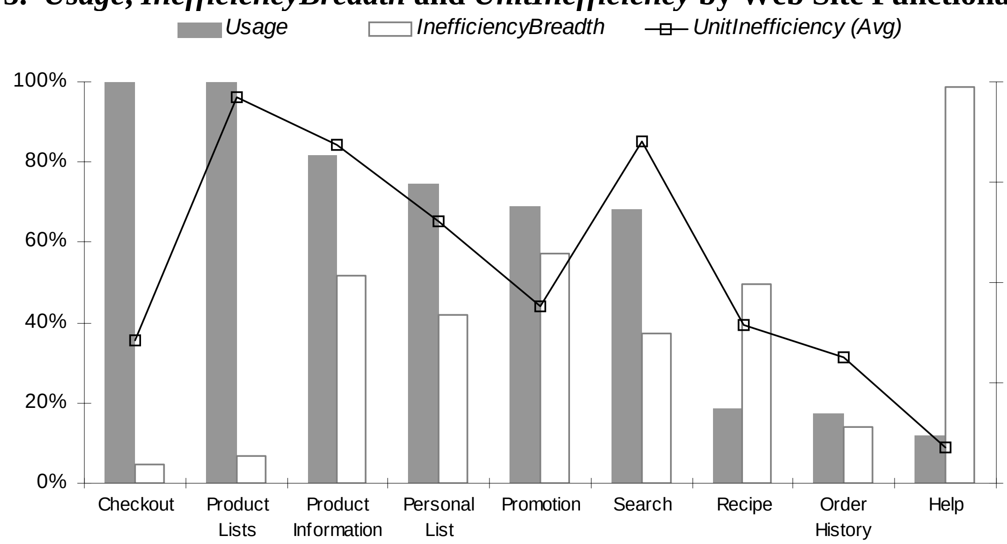
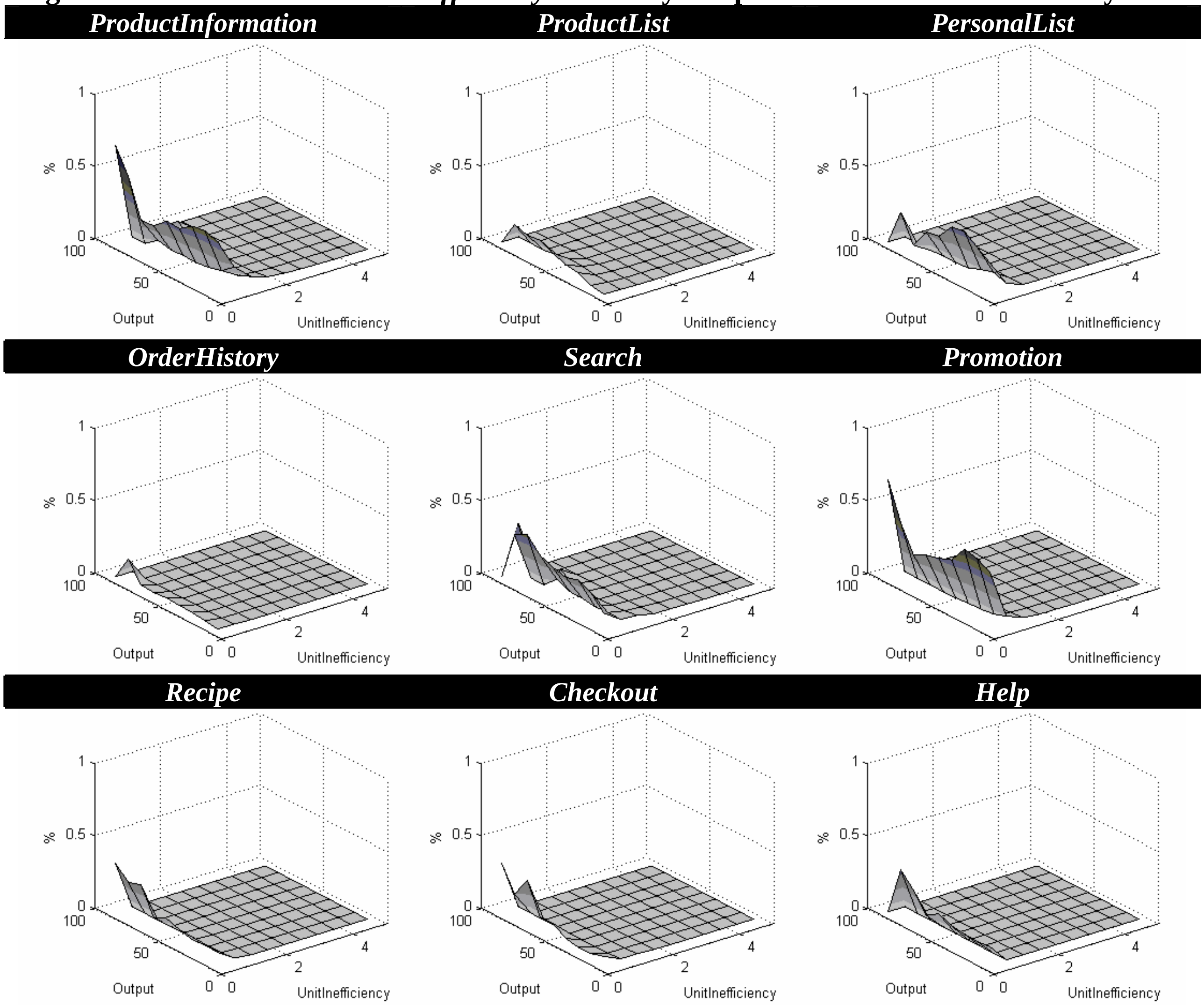


Figure 5. *Usage, InefficiencyBreadth* and *UnitInefficiency* by Web Site Functionality



Note: For clearer exposition, this graph uses two separate scales, with the proportional measures (i.e., *Usage* and *InefficiencyBreadth*) using the scale on the left-hand side, and average *UnitInefficiency* on the right-hand side. The Web site functionalities on the horizontal axis (x-axis) are sorted in decreasing order of usage rate (J_i / J).

Figure 6. Distribution of *UnitInefficiency* Scores by Output Volume and 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.