RANDOLPH E. BUCKLIN and CATARINA SISMEIRO*

Using the clickstream data recorded in Web server log files, the authors develop and estimate a model of the browsing behavior of visitors to a Web site. Two basic aspects of browsing behavior are examined: (1) the visitor's decisions to continue browsing (by submitting an additional page request) or to exit the site and (2) the length of time spent viewing each page. The authors propose a type II tobit model that captures both aspects of browsing behavior and handles the limitations of server log-file data. The authors fit the model to the individual-level browsing decisions of a random sample of 5000 visitors to the Web site of an Internet automotive company. Empirical results show that visitors' propensity to continue browsing changes dynamically as a function of the depth of a given site visit and the number of repeat visits to the site. The dynamics are consistent both with "within-site lock-in" or site "stickiness" and with learning that carries over repeat visits. In particular, repeat visits lead to reduced page-view propensities but not to reduced page-view durations. The results also reveal browsing patterns that may reflect visitors' time-saving strategies. Finally, the authors report that simple site metrics computed at the aggregate level diverge substantially from individual-level modeling results, which indicates the need for Web site analyses to control for cross-sectional heterogeneity.

A Model of Web Site Browsing Behavior Estimated on Clickstream Data

Since the commercial inception of the Internet, the ability of Web sites to track the behavior of their visitors has been considered one of the most promising facets of the new medium. The detailed records of Web usage behavior (clickstream data) provide researchers and practitioners with the opportunity to study how users browse or navigate Web sites and to assess site performance in various ways. Therefore, the use of clickstream data to model visitors' usage of a specific site and how that usage may change with experience could produce important dividends for researchers and practitioners interested in Web site design, Web site customization, and ongoing monitoring of a site's performance.

Despite the importance of analyzing how users browse a given site, to the best of our knowledge, no disaggregate clickstream model of detailed within-site browsing behavior and its potential dynamics has yet appeared in the marketing literature. In this article, we develop a modeling approach for understanding some of the basic aspects of within-site browsing behavior at the individual level. We conceptualize user navigation through an Internet Web site as a series of the following decisions: (1) whether to request an additional page (thereby remaining on the site) or to exit the site and (2) how long to view a given page on the site. We estimate a type II tobit model on the clickstream data collected by the operator of a major commercial Web site in the automotive industry. Our modeling results provide several findings about browsing behavior and a series of implications for Web managers.

The existing literature on Web site modeling can be classified into studies that analyze within-site behavior and studies that analyze across-site behavior. With respect to within-site browsing behavior, Huberman and colleagues (1998) propose a "law of surfing." In their model, the distribution of the number of pages requested by users could be accurately predicted by simple assumptions of surfing behavior. In particular, they assume that the utility of an additional page view is equal to the utility of the current page plus a normally distributed error. This produces an inverse Gaussian distribution for the predicted distribution of the number of

*Randolph E. Bucklin is Professor of Marketing, Anderson School, University of California, Los Angeles (e-mail: rbucklin@anderson.ucla.edu). Catarina Sismeiro is Assistant Professor of Marketing, Marshall School of Business, University of Southern California (e-mail: sismeiro@marshall.usc.edu). The authors thank an anonymous collaborating company for providing the data used in this study. Research support from the University of California, Los Angeles, Academic Senate is gratefully acknowledged. The authors thank three anonymous *JMR* reviewers, David Bell, Christophe Van del Bulte, and seminar participants at the Wharton School and Yale School of Organization and Management for comments.

Journal of Marketing Research
Vol. XL (August 2003), 249–267

page requests across visitors. The approach provides a good fit to Huberman and colleagues' data, but it does not incorporate the potential effect of covariates on page requests or take into account how the browsing behavior of site visitors may change as they browse the site or return to the site.

In an experimental study, Mandel and Johnson (2002) show that changes to Web page design affect the choices Web users make. They find that users dynamically adapt their behavior in response to the page-by-page stimuli presented to them, even if they are not aware of their own adaptive behavior. In light of Mandel and Johnson's results, we hypothesize that site visitors can change the way they browse a site as they request and view additional pages. For example, a change in browsing behavior can be a response to the salience of time constraints, which may vary as a visitor browses the site or may be the result of varying degrees of involvement with the site or with tasks being performed.

Johnson, Bellman, and Lohse (2003) study the duration of Web site sessions across multiple visits. They propose that the cognitive costs of using a site decrease with experience and that this can be modeled with the power law of practice, a simple functional form from cognitive science. Using session durations from a panel of Internet users, Johnson, Bellman, and Lohse test this learning phenomenon. Results indicate that visitors spend less time per session the more they visit the site, which in turn suggests that they become more efficient as they return to the site. Nevertheless, this model does not control for factors that might influence session duration across visits in the absence of learning effects. These factors include content of pages requested, server response time, and tasks performed during each site visit. Johnson, Bellman, and Lohse also do not address how visitors become more efficient; for example, returning visitors might request fewer pages per session because they already know where to go in the site, or they might view each page for a shorter amount of time because they complete tasks and process information more quickly. The disentanglement of these effects across multiple sessions is an objective of the modeling approach we propose.

Other marketing studies have investigated Web usage behavior across multiple Web sites and visits. In a laboratory experiment, Zauberman (2003) demonstrates that users quickly develop a tendency to become loyal to a single Web site to accomplish an assigned task. This occurred even when better Web sites quite literally were just a click away. Zauberman introduces the term "lock-in" to describe the behavior of subjects who he found unwilling to switch to other sites. Quantitative evidence for high levels of site loyalty comes from a study Johnson and colleagues (2002) conduct on Web user logs from a Media Metrix panel. They find that panelists engage in limited across-site search activities but are seldom strictly loyal to a single site in a category.

The evidence for site lock-in and limited across-site search behavior may be the result of a user's learning process that "spills over" multiple visits to the same site, similar to the process Johnson, Bellman, and Lohse (2003) suggest. When Internet users have learned how to use a site and can use the knowledge they acquired in one visit for subsequent visits, they may be reluctant to invest their time in learning how to use new sites. Our proposed modeling approach enables us to investigate how within-site browsing behavior changes as users return to a site.

Moe and Fader (2002a, b) report additional modeling research of interest. They use stochastic modeling approaches to investigate the repeat-visit behavior and purchase conversion rates of Web users at Amazon.com and CDnow. They find strong evidence of heterogeneity in Web usage behavior and changes in behavior over time and note the risks of ignoring these phenomena in evaluations of Web site performance. Moe and Fader (2002b) conclude that a potentially important area not yet investigated is the activity that takes place within a site visit.

In our modeling approach, we examine two aspects of within-site browsing behavior: (1) a visitor's series of binary decisions to stay or exit the site (manifest by page requests) and (2) the duration of each page view. Together, the two model components (page request and page-view duration) provide a parsimonious representation of the browsing decisions users face in a site visit. We jointly model page-request and page-view-duration decisions with a type II tobit model. Such a model is well suited for handling the censoring in log-file duration data, integrates both browsing decisions, and allows for a positive or negative correlation of the duration and page-request decisions.

We model each browsing decision as a function of user and site covariates. We focus on the effects that visit depth and repeat visitation have on users' staying with the site and on page-view duration. These two covariates enable the model to capture changes in browsing behavior that occur within and across site visits. We also include a series of additional covariates designed to control for content of the pages viewed, completion of certain tasks, and other site-specific factors. We show how to estimate our model using data recorded in the server log files from a commercial Web site. In light of the potential importance of heterogeneity in our data, we adopt a hierarchical Bayes approach and estimate the models using Markov chain Monte Carlo (MCMC) methods (e.g., Ansari, Essegaier, and Kohli 2000; Ansari and Mela 2002).

Our results indicate that Web site users dynamically adjust their browsing behavior both within and across visits. Among our findings, there first is evidence for within-site lock-in effects in both browsing components (page requests and page-view duration) as users go deeper into the site. Second, our results are consistent with the presence of learning effects that spill over multiple visits as users reduce their expected number of page views when they return to the site. For managers, we also show that aggregate-level statistics can be potentially misleading Web metrics.

This article is organized as follows: First, we describe the characteristics of log-file data, explain how we processed the data, and present the implications for modeling Web site browsing behavior. Second, we provide an overview of our approach to modeling browsing behavior and present the empirical results for the type II tobit model. Third, we discuss our results and the behavioral and Web site management implications of our findings. Fourth, we conclude by noting some of the limitations in our research and possible steps for further research.

SERVER LOG-FILE DATA

Our data set comprises the full server log files for October 1999 from a major commercial Web site in the automotive industry. The site permits visitors to view general company information and instructions on how to use the Web site, to research the configuration of almost all commercially available cars and light trucks, to obtain fixed "no-haggle" prices, and to place an order for a vehicle (visitors complete the remainder of the transaction by telephone, e-mail, or fax). The site is simple and linear and has standard search capabilities.

The log files register all the requests and information transferred between the client (the visitor's computer) and the company's commercial Web site server. Our log-file data are in the World Wide Web Consortium extended log-file format, which provides a rich set of information on every request made to the company's servers. Some of the fields stored are users' Internet protocol address and cookie content, browser and system used, number of bytes transmitted, server response time, time and day of request, and referring site. Given the time-sensitive content of most pages (e.g., pricing information and model availability that could potentially change at any time), the site forces an automatic request of pages from the server even if visitors use the back button. Visitors can access very few pages, mostly those about the contact and company information, using the back button with no additional request to the server.

Despite the richness of these log files, obtaining a meaningful data set required several important preprocessing tasks. For example, because each page can be formed by multiple frames and images, it was necessary to aggregate the requests that correspond to the same page. This ensured that our unit of analysis, page views, had meaning for users. In addition, we identified reloads, transmission errors, bytes transmitted, and server response time corresponding to each page requested (and not simply "hits"). Finally, we eliminated requests from Web crawlers, Web spiders, and the company's own administration personnel.

Additional processing was required to determine unique visitors and their multiple sessions. Because cookie identification information adequately approximates unique site users (see, e.g., Drèze and Zufryden 1998), we retained only those records with cookies (these constitute approximately 90% of the full data set). We then used cookies to identify unique visitors. We also defined a site visit as including two or more page views, and therefore we removed single-page visits from further analysis. If we were to retain visits with single-page requests, they would dominate the entire sample. Because (1) single page-view visits do not constitute "browsing" of the site in question and (2) the collaborating company had no practical reason for analyzing such visits, we retained only those visits with more than one page view. Management believed that most of the one-page visits were user mistakes, because almost all of the single page views were from site visitors who came to the site and never returned (more than 99.9%).

We followed previous research and assumed that a page request started a new session if it was requested after an idle period of at least 30 minutes (Catledge and Pitkow 1995). We then computed page-view duration as the time difference (in seconds) between consecutive page requests within a session.[1] Although we distinguished different sessions for the same user, we had no knowledge of the duration of the last page view for each session. This is a problem of censored data: We observed the exogenous variables for the last page request, but we did not observe the duration of the last page view of each session (see Amemiya 1985, pp. 360–61). This is because the company's own server log files had no way to record when a user moved to another site or closed the browser. Consequently, the total duration for each site visit was unknown.

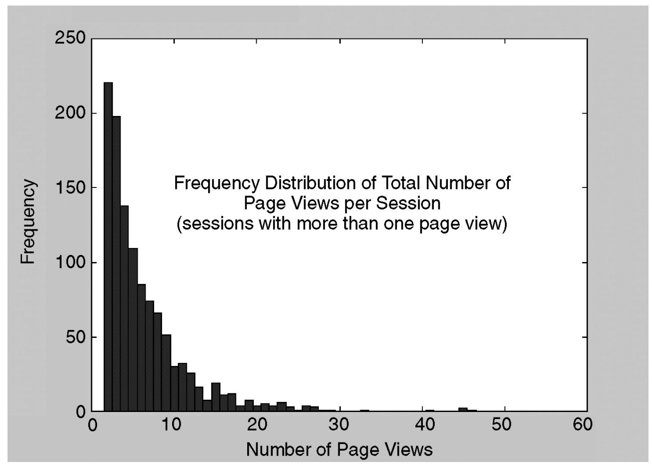
Table 1 presents summary statistics for the entire data set. Following our definition of a site visit, the data in Table 1 pertain to visitors who requested two or more page views. In Figure 1, we show the frequency distribution for the number of page requests by site visit. (Note that the distribution begins with two page views.) The mean number of pages per session is 5.95, and the mean number of sessions or visits per visitor is 1.45.

Table 2 presents the distribution of intervisit times for all repeat visitors. Users made virtually all repeat visits (97.6%) within 15 days of the previous visit. To obtain a manageable estimation sample, we randomly selected 5000 visitors (unique cookie identifications) from the second half of October 1999. This provided a reasonable two-week initialization period (first half of October) in which we captured previous site visits and studied across-visit dynamics. The two-week span also minimized the impact of minor site

Table 1
SUMMARY STATISTICS FOR BROWSING

	October 1999	*Random Sample, October 15–31*
Number of visitors	169,910	5000
Number of sessions	245,782	6630
Number of requests	1,462,457	40,560
Mean sessions per visitor	1.447	1.326
Mean pages per visitor	8.607	8.112
Mean pages per session	5.950	6.118
Mean page-view duration (seconds)	102.692	117.156

Figure 1
DISTRIBUTION OF THE TOTAL NUMBER OF PAGE VIEWS PER SESSION (SESSIONS WITH MORE THAN ONE PAGE VIEW)



[1]Page-view-duration time includes the download time and the time required to "assemble" the page on the client side (the site visitor's browser). Consequently, there is the need to control for differences in download and page assembly time among users and pages.

Table 2

DISTRIBUTION OF INTERVISIT TIME FOR REPEAT VISITS (OCTOBER 1999)

Days	*Visits*	*Percentage of Total Visits*	*Cumulative Visits*	*Cumulative Percentage of Total Visits*
0	34,575	45.6	34,575	45.6
1	10,917	14.4	45,492	60.0
2	6575	8.7	52,067	68.6
3	4712	6.2	56,779	74.8
4	3489	4.6	60,268	79.4
5	2802	3.7	63,070	83.1
6	2637	3.5	65,707	86.6
7	2041	2.7	67,748	89.3
8	1376	1.8	69,124	91.1
9	1019	1.3	70,143	92.5
10	917	1.2	71,060	93.7
11	783	1.0	71,843	94.7
12	680	.9	72,523	95.6
13	632	.8	73,155	96.4
14	544	.7	73,699	97.1
15	381	.5	74,080	97.6
>15	1791	2.4	75,871	100.0

Notes: Intervisit times of 0 days indicate that the visitor returned to the site in less than 24 hours; 1 day indicates that the visitor returned to the site in 24 to 48 hours, and so on.

Table 3

DISTRIBUTION OF THE NUMBER OF SESSIONS PER VISITOR FOR THE SAMPLE

Number of Sessions	*Percentage of Total Visitors*	*Cumulative Percentages*	*Remaining Visitors (Percentages)*
1	79.3	79.3	20.7
2	11.8	91.1	8.9
3	4.1	95.3	4.8
4	1.9	97.2	2.8
5	1.0	98.2	1.8
6	.6	98.8	1.2
>6	.4	99.1	.9

changes that might have taken place. (No major site change took place during this period, and we are unaware of specific changes to the page links or content.) Although we used a subsample of 15 days, we retained the behavior data of each selected individual for the entire month of October. The final sample comprised 6630 sessions that corresponded to 40,560 page requests. Of these requests, 33,930 were made within the site to view an additional page and 6630 were exiting requests (one for each session). Approximately 21% of visitors in the sample made two or more visits to the site during the period considered (for the distribution of the number of sessions per visitor from the sample, see Table 3). Finally, we found no significant differences between the descriptive statistics for the full sample and our random subsample (Table 1).

MODELING APPROACH

The objective of our model is to capture browsing behavior for a single Web site, both within and across visits, and to do so at the disaggregate level of individual user behavior. We selected two aspects of disaggregate browsing behavior that jointly provided a parsimonious yet complete view of the decisions users face during each site visit. The first aspect is the user's decision of whether to request an additional page from the site (i.e., a stay/exit decision). Collectively, this series of decisions determines the total number of pages requested by a visitor for a given session. The second aspect is the user's decision of how long to view each page, given the stay/exit decision. The combination of stay/exit and page-view-duration decisions determines the total duration of each visitor's session.

In focusing on a visitor's decision of whether to request an additional page and how long to spend viewing it, we limited our study to those Internet users who had already chosen to go to the Web site and begin to navigate it. We did not model the choice of site or the browsing behavior of users who visited the Web site, requested a single page, and then exited (for these users, there is in effect no "browsing" history). We therefore defined a visit to a Web site as comprising two or more page requests. In practical terms, this means that in our analysis we did not consider visitors who accessed only the site's home page.

Other dependent variables can also describe the within-site navigation behavior of Internet users. For example, a visitor's navigation decisions can be conceptualized as following a tree-shaped path through the site. In this study, we did not consider users' specific page choices (e.g., the choice among links presented on a current page), leaving the issue of specific link choice for further research. In addition, we did not explicitly model users' decisions of whether to return to the Web site, because our focus is on browsing behavior within a given site visit. Nevertheless, we included covariates in our model that capture previous site visitation and the depth of each site visit; these enabled us to analyze the dynamics of within-site behavior both within and across visits.

Type II Tobit

We jointly modeled the page-request and page-view-duration decisions with Heckman's (1978) generalized tobit model (type II tobit). Within the marketing field, DeSarbo and Choi (1999) provide a search-behavior application of this model, and DeSarbo and Jedidi (1995) present a consideration-set application. A type II tobit model provides two appealing features.

First, it is well suited for the type of censoring typical in log-file duration data. Log files register only the page requests visitors make to a company's servers. Page-view durations are then computed from the time differences (in seconds) between consecutive page requests. Because log files do not record when a visitor closed the browser or decided to visit another Web site (those actions do not correspond to requests to the company's servers), we cannot determine the page-view duration of the last page of any session. For example, split hazard models cannot handle log-file data because the page-view duration of the last page in each visit is not observed at all and does not have any logical upper or lower bound. In principle, a censored survival model could be used to model page requests and durations, but it would impose the restriction that the same parameters govern the page-request and page-view-duration decisions. We elected to use the type II tobit to allow for differences in these two decision processes.

The typical application of type II tobit models (e.g., in the area of labor economics and search behavior) is different from that presented here, despite an identical likelihood function specification. The type of censoring observed with log-file data is the result of the data collection process. There is no behavioral reason we cannot observe the duration of the last page view: Such duration exists, but there simply is no measure for it. In contrast, for most search applications of the type II tobit model, there are behavioral reasons for not observing certain values of the continuous dependent variable. For example, workers may decide not to participate in the labor market, and as a result zero hours of employment are observed; consumers may decide they do not need to search, and consequently no search is observed. In the area of selectivity bias in choice and quantity decisions, consumers may decide not to buy anything, and consequently no quantity purchased or no dollars spent is observed. Many of these applications correspond to double-hurdle models in which the continuous variable is observed only if a latent construct is positive. In our case, page-view durations by definition are positive, and they all occur; we simply cannot measure all of them.

Second, an appealing feature of the type II tobit is that it integrates page-view and page-request decisions while allowing for either a positive or a negative correlation of both browsing components. By jointly modeling the two decisions and recognizing the possible selectivity bias in page-view duration, we can test whether page-request decisions correlate with page-view durations, and vice versa. For example, greater page-request propensities might be associated with longer or shorter page-view durations. In either case, we need to account for such potential correlation, and the type II tobit model can do this.

With a disaggregate-level model, we can account for and study the simultaneous impact of multiple covariates, and we can handle visitor heterogeneity. In doing so, we avoid the potentially misleading results associated with aggregate-level statistics monitored by Web managers, such as average number of page views and their average duration. These aggregate-level statistics can confound cross-sectional and longitudinal effects, whereas an individual-level model of browsing can distinguish these effects.

The dependent variables in the model are directly relevant to Web managers who monitor site performance. Two measures managers track are average number of pages requested and average session duration (from which average page-view duration is computed). The monitoring of these metrics, together with monitoring the number of unique visitors, is also an essential tool for the management of Web server capacity. As a result, it is important to understand what determines visitors' page-request decisions and their decisions of how long to spend viewing each page. By modeling page requests and page-view duration, we can better understand how page-specific covariates influence site visitors' behavior within and across visits and how visitors dynamically adjust their behavior. Such tasks could not be accomplished had we chosen the number of pages requested in a visit and the total visit duration as dependent variables.

The type II tobit model includes an index equation (additional page-request versus site-exit decision), a structural equation (page-view-duration decision), a threshold equation that links the page-view decision and the observed duration, and a stochastic specification. We subsequently describe each component, present the likelihood function, and introduce the random-effects approach to heterogeneity.

Page-Request Decision

We model a visitor's decision to make an additional page request or exit the site each time the visitor has an opportunity to request an additional page view. Our model therefore predicts the probability of an additional page request occurring, given that the visitor has arrived at the current page view. We use a hierarchical binary probit model, which enables us to estimate visitor-specific parameters for the covariates.

As noted previously, Huberman and colleagues (1998) model the number of links that a user follows in a Web site, which is a related but different dependent variable. They assume that there is value in each page visited and that clicking on the next page reveals that it is valuable as well. Because the value of the next page is uncertain, Huberman and colleagues assume that it is stochastically related to the current one. In other words, the value of the next page view is the value of the current page view plus or minus a random term. Users continue browsing until they perceive the expected cost of continuing as greater than the discounted expected value of the information to be found in the future. From these assumptions, Huberman and colleagues derive the probability of the number of links that a user follows in a Web site.

As do Huberman and colleagues (1998), we assume that Web users continue proceeding to additional pages as long as the value of the next page, or its utility, exceeds some threshold. We also assume that the utility of the next page is not certain but is stochastically related to the value of current and past pages. In other words, the visitor uses current and past page characteristics to infer the utility of an additional page request versus the utility of a site exit. In contrast with Huberman and colleagues, we use covariates to model the "attractiveness" of an additional page request at the disaggregate level. By modeling the choice decisions pertaining to each page request (rather than predicting the number of pages viewed), we can better understand the dynamics that characterize within-site browsing behavior and the potential causal impact of relevant covariates.

We now turn to the formulation of our binary probit choice model and define page-view utility as

$$y^*_{1ij} = \mathbf{x}_{1ij}\beta_{1i} + u_{1ij}, \tag{1}$$

where y^*_{1ij} is the utility associated with the request of an additional page within the site for visitor i = 1, …, N, and page-view occasion j = 1, …, J_i (J_i is the number of page views of visitor i). The vector β_{1i} is a ($p_1 \times 1$) vector of visitor-specific parameters, $\mathbf{x}_{1ij}$ is a ($1 \times p_1$) vector of individual- and page-specific covariates that includes an intercept, and u_{1ij} is a normally distributed error term. To simplify the presentation of the model, we have suppressed visit-specific subscripts. The page-specific subscript, j, may cross multiple visits.

We observed only page requests and site exits, not utilities. Thus, we assume that a positive utility is associated with an additional page view and a negative page-request utility is associated with a site exit. If we code these events as $\delta_{ij} = 1$ if there was an additional page request and $\delta_{ij} = 0$ if there was a site exit (no additional page requests), then we observe δ_{ij} such that

$$\delta_{ij} = 1 \text{ if } y^*_{1ij} \geq 0, \text{ and } \delta_{ij} = 0 \text{ if } y^*_{1ij} < 0. \tag{2}$$

Page-View-Duration Decision

We also modeled the duration of a specific individual page view as a function of user- and page-specific covariates. For example, the covariates account for page characteristics, time required for data transmission, user activity, and completion of specific tasks. We modeled the logarithm of the page-view duration, y^*_{2ij}, as a linear function of the covariates such that

$$y^*_{2ij} = \mathbf{x}_{2ij}\beta_{2i} + u_{2ij}, \tag{3}$$

where y^*_{2ij} is the log of the page-view duration for visitor i = 1, …, N, and page-view occasion j = 1, …, J_i. The vector β_{2i} is a ($p_2 \times 1$) vector of visitor-specific parameters, $\mathbf{x}_{2ij}$ is a ($1 \times p_2$) vector of covariates that includes an intercept, and u_{2ij} is a normally distributed error term. Because of the data structure of log files, we observed only the duration of page views that are followed by additional page requests. We did not observe the duration of the last page view in each session. Consequently, we define y_{2ij} as the observed log of the page-view duration for individual i and page request j (j = 1, …, J_i, and i = 1, …, N), such that

$$y_{2ij} = y^*_{2ij} \text{ if } \delta_{ij} = 1, \text{ and } y_{2ij} \text{ is not observed for } \delta_{ij} = 0. \tag{4}$$

Note that we specified a log-normal model instead of a normal model for page-view duration. Historical applications of type II tobit models usually have studied behavioral outcomes that occur only when positive. In such cases, y^*_{2ij} is only observed when positive: $y_{2ij} = y^*_{2ij}$ if and only if $y^*_{2ij} >$ 0 and $\delta_{ij} = 1$; $y_{2ij} = 0$ otherwise. Such a behavioral structure corresponds to a double-hurdle model specification, which does not fit the phenomenon we model. In the case of page-view durations, although we did not observe the duration of the last page view, these durations actually exist and are also positive. Therefore, a normal linear model would be inappropriate because it allows for negative duration predictions even though all page-view durations, observed and unobserved, can only be positive. Consequently, we log-transformed the page-view duration and adopted a log-normal specification.[2]

Error Structure and Likelihood

In our specification of the type II tobit model, we assumed that the bivariate joint random error, $(u_{1ij}, u_{2ij})'$, was identically and independently distributed across pages. We captured differences and dependences among pages, within and across site sessions, through page- and session-specific covariates. We further assumed that the error term is bivariate normal such that

$$\begin{bmatrix} u_{1ij} \\ u_{2ij} \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right), \tag{5}$$

where σ_1^2 is the variance of the utility error terms, σ_2^2 is the variance of the error terms for log of page-view duration, and σ_{12} is the covariance between the error terms of the two model components. For identification purposes, we set σ_1^2 equal to 1. Such a constraint identifies the choice and duration parameters and is the one typically employed in probit analysis. With the full specification of the error structure, we can derive the likelihood function for each individual i = 1, …, N (see Amemiya 1985, pp. 385–87):

$$L_i = \prod_{j=1}^{J_i} \left[1 - \Phi(\mathbf{x}_{1ij}\beta_{1i})\right]^{(1-\delta_{ij})} \times \left\{\Phi\left[\mathbf{x}_{1ij}\beta_{1i} + \rho\sigma_2^{-1}(y_{2ij} - \mathbf{x}_{2ij}\beta_{2i})\,(1-\rho^2)^{-\frac{1}{2}}\right] \phi\left(\frac{y_{2ij} - \mathbf{x}_{2ij}\beta_{2i}}{\sigma_2}\right)\Big/\sigma_2\right\}^{\delta_{ij}}, \tag{6}$$

where ρ is the correlation between the error terms of the two model components u_{1ij} and u_{2ij} (i.e., $\rho = \sigma_{12}/[\sigma_2\sigma_1] = \sigma_{12}/\sigma_2$), $\Phi(\cdot)$ represents the standard normal cumulative distribution function, and $\phi(\cdot)$ represents the standard normal probability density function. To test for the independence of the two decision components, we simply tested for $\rho \neq 0$. If ρ is positive, longer page-view durations are associated with higher propensity for additional page requests; if ρ is negative, longer page-view durations are associated with lower propensity for additional page requests (i.e., site visits that have fewer page views on average).

Heterogeneity

To model heterogeneity across visitors, we adopted a flexible random coefficient model for both the probit and the duration components. Random-coefficients tobit models

[2]We tested for alternative page-view-duration model formulations. Within the type II tobit framework, we tested the log-normal specification against a simple normal model. Although the effects of all covariates were in the same direction, the log-normal model outperformed the normal formulation. We also in- and out-of-sample tested the type II tobit log-normal model against a Weibull specification of page-view duration. To accommodate interdependent decision components, we allowed a polynomial function of the page-request expected utility to affect the Weibull hazard. The type II tobit model outperformed this alternative formulation, but the substantive results remained identical. Details are available from the authors on request.

have been previously studied in the literature. For example, Kamakura and Wedel (2001) propose an exploratory tobit factor analysis that imposes a factor structure on the covariances, which makes the model feasible for large multivariate problems. For each covariate k, and given all the other covariates (denoted by –k),

$$(7)\quad \beta_{hki}\left|\beta_{h-ki} \sim \mathbf{N}\left(m_{hk}, s_{hk}^2\right)\right. \text{ for } k = 1, \ldots, N, \text{ and } h = 1, 2,$$

where m_{hk} is the mean for covariate k in the page-request model (h = 1) or the duration model (h = 2), and s_{hk}^2 is the corresponding variance. Thus, our approach assumed conditional independence for the duration and page-request model parameters. (We note that a full posterior covariance matrix is defined and can be computed from the MCMC draws obtained from the sampling of the posterior distribution.) We also assumed conditional independence across observations, both within and across sessions.

A random-effects approach, which stochastically pools data across users (Heckman and Singer 1984), is frequently employed in the analysis of choice models using panel data. It has been extensively applied in both marketing and economic studies (for a review, see Allenby and Rossi 1999; Wedel et al. 1999). We believe that this is an appropriate model of heterogeneity, given our data. The large number of subjects in our data set makes a fixed-effect or a discrete approach to heterogeneity less appealing because the potential number of mass points for the parameters increases with the number of users. For example, if model estimation results in too many visitor segments, interpretability of results becomes difficult.

EMPIRICAL ANALYSIS

We applied the type II tobit model to the Web page browsing behavior of our sample of Web site visitors. We now describe the covariates, discuss how we operationalized them, and present our model results.

Model Variables

To capture the dynamics of the stay/exit and duration decisions, both within and across visits, we modeled each decision as a function of visit depth and repeat visits. We also identified a series of additional covariates that we believed might affect these decisions and therefore might be important to hold constant when studying the effects of visit depth and repeat visits.

Visit depth. As our measure of visit depth, we used the cumulative number of session page views it takes visitor i to arrive at page j, or $CPAGE_{ij}$. If there were time constraints on Internet usage at the individual level, we expected visit depth to have negative effects on the probability of an additional page request and on expected page-view duration. In contrast, if visitors became more involved as they requested more pages, we expected visit depth to be associated with higher probabilities of page requests and longer page-view durations. This could be considered within-site lock-in, analogous to Zauberman's (2003) across-site lock-in finding. These two competing hypotheses (time constraints versus involvement or lock-in) lead to the alternate possibilities displayed on the left-hand side of Figure 2.

Repeat visits. As our measure of repeat visits, we used the cumulative number of site visits made by visitor i at page view j, or $CSESSION_{ij}$. Repeat visits enabled us to test for learning effects as experience with the site increases. For example, if users are able to learn how to navigate the site efficiently and if they do not forget acquired knowledge from previous visits in subsequent site visits, they may be able to go to the pages of interest more quickly and may need less time at each page to perform the intended tasks. Johnson, Bellman, and Lohse (2003) provide evidence from e-commerce sites in the books, music, and travel categories that this learning phenomenon occurs. Their results indicate that visitors spend less time per session the more they visit the site, but there is no indication of whether reduced session duration is due to fewer page requests or shorter page-view durations.

The opposite effects could be expected for repeat visits if users who returned to the site were more involved with the browsing process and became more careful. Then, the probability of page requests and the average page-view duration would increase with an increase in repeat visits. The right-hand side of Figure 2 presents the possible patterns of browsing behavior dynamics with respect to the number of repeat visits.

Additional covariates. Browsing behavior is also affected by many factors other than visit depth and repeat visits. For example, Dellaert and Kahn (1999) show that longer waiting periods during page downloads can negatively affect visitors' evaluations of Web sites. Zauberman (2003) demonstrates that requiring extra input effort from users (in the form of higher setup costs) can reduce satisfaction with the entire search experience. Nielsen's (2000) studies reveal that usable Web sites place a premium on error prevention and that users show low tolerance for any problems. In addition, because page-view durations include download and assembly time, there is the need to control for these factors. Using the available log-file data, we developed an additional set of covariates designed to capture factors such as download effort, interaction between user and Web site, tasks performed by the user, system reliability, and system response time.[3] Appendix A gives a short description of each of these variables; Table 4 presents the corresponding summary statistics.

Estimation

To estimate the hierarchical type II tobit model of page requests and page-view durations, we used a Bayesian approach implemented with MCMC methods. To complete our model specification, we introduced priors over the parameters common to all visitors (see Appendix B).

We sampled from the joint posterior distribution by sampling from the full conditional distributions of the model parameters. We used data augmentation methods to simulate the latent utilities conditional on the model parameters (Tanner and Wong 1987). We also applied standard theory for the normal hierarchical model (Gelfand et al. 1990; Gelfand and Smith 1990) to obtain the full posteriors of m_{1k}, s_{1k}^2, m_{2r}, s_{2r}^2

[3]We note that the data of our empirical analysis are from October 1999. During that time, the company management believed that few visitors were broadband users, which posed few problems with respect to differences in download time across users. The individual-level intercepts and parameters for the covariates related to download time (BYTES, SRVRESP, and DYNAMIC) account for the remaining differences among users' connection speeds.

Figure 2
HYPOTHESES FOR DYNAMIC BROWSING BEHAVIOR

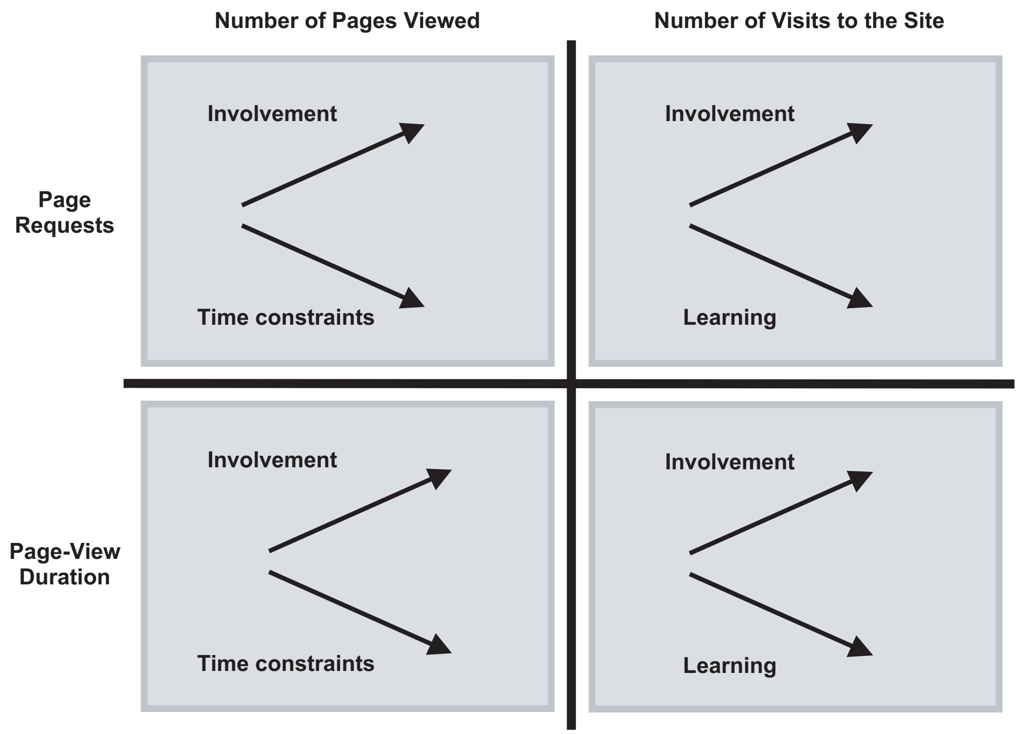


Table 4
SUMMARY STATISTICS FOR COVARIATES

Variable Name and Abbreviation	*Mean*	*Standard Deviation*
Bytes transferred (BYTES)	.275	.238
Number of cars configured (CMAKE)	.201	.817
Visit depth (CPAGE)	5.886	6.186
Repeat visits (CSESSION)	1.921	2.208
Dynamic page content (DYNAMIC)	.126	.332
Downloading error (ERROR)	.081	.273
First car already configured (FIRSTMAKE)	.146	.354
Car previously ordered (ORDER)	.005	.069
Previous page-view duration (PREVDUR)	1.633	2.909
Page reloaded (RELOAD)	.233	.423
Server response time (SRVRESP)	.003	.041

(k = 1, …, p_1, and r = 1, …, p_2), and β_{2ir} (i = 1, …, N, and r = 1, …, p_2), conditional on the remaining model parameters. These are of known functional forms.

When estimating independent page-request and duration models, standard theory for the normal hierarchical model also applies for the probit individual-level parameters, β_{1ik} (i = 1, …, N, and k = 1, …, p_1) and for the log-duration variance, σ_2^2. However, when estimating the full type II tobit model, the conditional distributions of these parameters and the conditional distribution for the correlation term ρ are not of any known form. We sampled from the corresponding full conditionals using Gilks and Wild's (1992) adaptive rejection sampling (ARS) algorithm if the full conditionals were log-concave (see also Gilks 1992) and Gilks, Best, and Tan's (1995) adaptive rejection metropolis sampling (ARMS) algorithm if the conditionals were not log-concave. Appendix C gives the full conditional distributions and the details on the Gibbs sampler, data augmentation procedures, and ARS and ARMS steps.

Model Selection

Bayes factors are widely used to make comparisons among statistical models; however, given the complexity of our model, Bayes factors are computationally nontrivial. We therefore followed Gelfand (1996) and used the cross-validation predictive densities and pseudo–Bayes factor (PSBF) to perform model selection (for details, see Appendix D); PSBFs have been used in marketing for model and variable selection (DeSarbo, Kim, and Fong 1999; DeSarbo et al. 1998) and are widely used in both economics and life sciences literature (e.g., Dey, Chang, and Ray 1996; Sahu et al. 1997).

We tested linear and nonlinear (logarithmic and quadratic) functional forms for the predictor variables and selected the specifications that provided the best PSBF. For both visit depth (CPAGE) and repeat visits (CSESSION), a log form provided the best fit. We performed the same analyses for the remaining variables.

We also tested additional variables not reported here; among them were cumulative bytes transferred, number of page reloads while transferring the current page, and cumulative number of hits as stored by the cookie. These additional variables did not provide model improvements when compared on PSBFs. Using a simple probit model that does not account for individual-level heterogeneity, we also tested for the presence of first-order autoregressive utilities. We found no evidence for first-order autoregression at the utility level. (Note that the test likely favors the presence of such autocorrelation, because the presence of unaccounted for heterogeneity exacerbates state dependence.)

We tested an alternative model formulation for the page-request decision by modeling the number of pages requested from the site by each visitor in each site visit, using a Weibull hazard specification with gamma heterogeneity. (We thank an anonymous reviewer for suggesting this alternative model.) We selected the best specification by sequentially adding covariates and testing their functional form using PSBFs. We then computed the corresponding discrete-time hazard to predict the probability of an additional page request at each page view and to determine the contribution for the log of PSBF. To allow for a fair comparison, we re-estimated the best probit specification with heterogeneity only in the intercept (we set response parameters to be common across all visitors) and computed the PSBF of the probit model with respect to hazard specification. We found that the probit model clearly outperformed the alternative hazard formulation: The contribution for the log PSBF is –16,819.1 for the probit and –25,514.7 for the alternative model.

For the duration component, we also tested a Weibull model with gamma heterogeneity (e.g., Jain and Vilcassim 1991). The substantive results for this alternative model were identical to the results presented here. The Weibull model also failed to provide any improvement in PSBF.

Finally, to test alternative heterogeneity formulations, we estimated latent class models for both the page-request and the page-view-duration decisions. For the page-request model, we obtained solutions up to and including six segments. (We could not obtain convergence for seven or more segments because of the large number of parameters and local optima problems.) Of the models estimated, the Bayesian information criteria selected the six-segment solution. In the six-segment solution, the parameters for CPAGE, CSESSION, CMAKE, PREVDUR, and SRVRESP have the same sign, across all segments, as those reported for our model. The control variables ERROR, RELOAD, and BYTES have differing signs across segments, but many of the parameters were not significant (see Appendix A). We acknowledge the possibility that more segments, if estimable, would improve model fit. For the duration model, a three-segment solution provided the best Bayesian information criteria. The three segments were similar, and the substantive results from all segments were identical to the type II tobit results (the only exception was the CMAKE parameter in one of the segments; however, this was a non-significant parameter). Collectively, these results indicate that our findings are largely robust to heterogeneity specification.

We performed a comparison of the different model formulations both in-sample and with a holdout sample of visitors. We compared model formulations using the PSBF, mean and median squared error, mean and median absolute deviation, and correlation of predictions and observed values. The duration model presented here outperformed all alternatives based on PSBF, median squared error, mean and median absolute deviation, and correlation of predictions and observed values. Only for the mean squared error did the model not outperform the alternative formulations, which can be explained by the need to exponentiate the predictions of log durations to obtain the duration variable, which exacerbates extreme error terms. When the comparison is based on the prediction of log duration, the log-normal type II tobit formulation is superior on all fit criteria.

The final set of covariates used in the best-fitting choice model utility is given by

$$\text{(8)} \qquad x_{1ij} = \begin{bmatrix} \text{INT, BYTES}_{ij}\text{, SRVRESP}_{ij}\text{,} \\ \text{ERROR}_{ij}\text{, RELOAD}_{ij}\text{, ln(CPAGE}_{ij}\text{),} \\ \text{ln(PREVDUR}_{ij} + 1\text{), ln(CSESSION}_{ij}\text{),} \\ \text{ln(CMAKE}_{ij} + 1\text{)} \end{bmatrix},$$

where INT corresponds to the intercept. Note that 1 is added to CMAKE and PREVDUR to ensure that they are greater than 0. For the duration model, the best-fitting specification is given by

$$\text{(9)} \quad x_{2ij} = \begin{bmatrix} \text{INT, BYTES}_{ij}\text{, SRVRESP}_{ij}\text{, ERROR}_{ij}\text{, RELOAD}_{ij}\text{,} \\ \text{ln(CPAGE}_{ij}\text{), DYNAMIC}_{ij}\text{, CMAKE}_{ij} \end{bmatrix}.$$

As we discuss subsequently, repeat visits (CSESSION) did not improve the fit of the duration model and are therefore omitted from Equation 9.

Tables 5 and 6 summarize the results from fitting the type II tobit models for various specifications of the covariates, ranging from intercept only to full specification. These include the binary probit (Table 5) and page-view-duration (Table 6) components. For the probit model, the log PSBF selects Model C9 as best fitting. For comparison, the fit of Model C9 without heterogeneity is presented alongside, demonstrating the importance of individual differences in page-request decisions.

For the duration model, the intercept-only model is presented first as Model D1. Models D1–D8 assume independence of choice and duration decisions, whereas Model D9 includes the correlation term between model errors. As is shown in Table 6, Model D8 is the best-fitting model for duration. The 95% probability intervals for ρ, from Model D9, are [–.014,.029], which means that, given the proposed model formulation, the two browsing components are independent. Thus, longer or shorter page-view durations are not associated with higher or lower page-request propensities (and vice versa). Therefore, both browsing components can be analyzed independently.

For fit comparison, we also estimated Model D8 without heterogeneity; results are presented in the last column of Table 6. A comparison of the results for Model D8 with and without heterogeneity demonstrates the importance of individual differences in page-view-duration decisions. We now discuss the parameter estimation results for each of the best-fitting models.

Results

Table 7 presents a summary of the population mean results (posterior means and 95% coverage intervals) for the parameters of the type II tobit model components: the binary probit and the duration regression. We note that the 95% coverage intervals for all predictor variable parameters cover the same sign as the population means. We discuss the results of the best-fitting models based on the log PSBF (see Table 6). For the choice and duration components, these are Models C9 and D8, respectively. In Table 8, we present the cross-sectional mean and range of the posterior means of the individual-level choice and duration elasticities. For dummy

Table 5
PAGE-REQUEST MODELS (MODEL FITS, POSTERIOR MEANS, AND 95% PROBABILITY INTERVALS)

	Model C1	*Model C2*	*Model C3*	*Model C4*	*Model C5*	*Model C6*	*Model C7*	*Model C8*	*Model C9*	*No Heterogeneity Model C9*
INTERCEPT	.960	.848	.867	.873	.861	1.090	1.753	1.878	1.885	1.242
	[.942,.977]	[.823,.870]	[.846,.889]	[.852,.894]	[.838,.887]	[1.063,1.119]	[1.689,1.810]	[1.819,1.935]	[1.834,1.943]	[1.190,1.1297]
BYTES		.420	.413	.540	.412	.373	.217	.203	.211	.260
		[.353,.497]	[.349,.484]	[.472,.600]	[.350,.491]	[.289,.448]	[.135,.309]	[.111,.307]	[.126,.308]	[.189,.332]
SRVRESP			–8.674	–8.807	–8.745	–7.842	–8.267	–8.646	–8.485	–1.476
			[–10.670,–7.050]	[–10.062,–7.478]	[–10.098,–7.518]	[–9.186,–6.753]	[–9.878,–6.846]	[–10.105,–7.330]	[–9.434,–7.445]	[–1.820,–1.137]
RELOAD				–.150	–.122	–.091	–.047	–.042	–.050	–.089
				[–.193,–.113]	[–.164,–.081]	[–.135,–.046]	[–.092,–.004]	[–.091,.005]	[–.102,–.006]	[–.123,–.054]
ERROR					.780	.692	.426	.450	.427	.459
					[.686,.870]	[.570,.788]	[.331,.523]	[.369,.545]	[.317,.542]	[.386,.532]
LN(PREVDUR + 1)						–.331	–.219	–.225	–.220	–.247
						[–.358,–.302]	[–.257,–.184]	[–.263,–.191]	[–.250,–.189]	[–.272.–.223]
LN(CPAGE)							–.632	–.693	–.711	–.087
							[–.679,–.574]	[–.743,–.650]	[–.755,–.668]	[–.104,–.071]
LN(CSESSION)								–.117	–.194	–.069
								[–.213,–.137]	[–.236,–.151]	[–.101,–.037]
LN(CMAKE + 1)									.186	.102
									[.105,.269]	[.056,.148]
Contributions to log PSBF	–18,111.6	–18,070.8	–17,869.4	–17,866.5	–17,711.3	–17,067.9	–16,864.9	–16,718.6	–16,703.8	17,453.6
Number of covariates	1	2	3	4	5	6	7	8	9	9
Log PSBF	—	40.8	242.2	245.1	400.3	1043.7	1246.7	1393.0	1407.8	658.0

Notes: The variables ORDER, DYNAMIC, and FIRSTMAKE did not provide any improvement of the PSBF; the Log PSFB obtained by adding each variable to Model C9 is 1385.1 for ORDER, 1407.7 for DYNAMIC, and 1259.9 for FIRSTMAKE.

Table 6
DURATION MODEL (MODEL FITS, POSTERIOR MEANS, AND 95% PROBABILITY INTERVALS)

	Model D1	*Model D2*	*Model D3*	*Model D4*	*Model D5*	*Model D6*	*Model D7*	*Model D8*	*Full Tobit Model D9*	*No Heterogeneity Model D9*
INTERCEPT	4.104	3.913	3.910	3.882	3.891	3.768	3.833	3.838	3.825	3.917
	[4.088,4.120]	[3.893,3.933]	[3.888,3.931]	[3.863,3.901]	[3.872,3.911]	[3.743,3.795]	[3.804,3.859]	[3.813,3.867]	[3.787,3.871]	[3.876,3.958]
BYTES		.660	.652	.393	.455	.523	.522	.558	.814	.419
		[.619,.710]	[.603,.703]	[.341,.439]	[.405,.507]	[.475,.565]	[.509,.605]	[.503,.613]	[.783,.858]	[.372,.466]
SRVRESP			7.870	8.562	8.423	7.607	8.549	8.016	14.411	1.620
			[5.943,9.952]	[6.853,10.324]	[6.589,10.330]	[5.642,10.047]	[6.117,10.925]	[5.855,10.272]	[13.664,15.184]	[1.271,1.987]
RELOAD				.435	.421	.392	.371	.373	.564	.456
				[.405,.463]	[.391,.450]	[.364,.423]	[.344,.399]	[.344,.404]	[.525,.606]	[.431,.482]
ERROR					.234	.171	.155	.155	.182	.164
					[.197,.268]	[.131,.212]	[.106,.203]	[.100,.216]	[.147,.210]	[.102,.230]
LN(CPAGE)						.111	.078	.079	.364	.013
						[.095,.126]	[.060,.095]	[.063,.095]	[.337,.405]	[.006,.022]
DYNAMIC							–.401	–.400	–.369	–.415
							[–.436,–.361]	[–.450,–.354]	[–.385,–.344]	[–.445,–.386]
CMAKE								–.062	–.064	.005
								[–.102,–.028]	[–.116,–.020]	[–.008,.018]
Contributions to log PSBF	–189,803.0	–189,427.3	–189,358.9	–188,931.9	–188,881.9	–188,800.3	–188,698.8	–188,680.5	–189,201.3	–189,030.8
Number of covariates	1	2	3	4	5	6	7	8	8	8
Log PSBF	—	375.7	441.1	871.1	921.1	1002.7	1104.2	1122.5	601.7	772.2

Notes: The variables CSESSION, ORDER, PREVDUR, and FIRSTMAKE did not provide any improvement of the PSBF; the Log PSBF obtained by adding each variable to Model D8 is 1072.6 for CSESSION, 1109.3 for LN(CESSION), 1117.3 for ORDER, 1090.0 for PREVDUR, and 1107.6 for FIRSTMAKE.

Table 7

MODEL PARAMETER ESTIMATES (POSTERIOR MEANS AND 95% PROBABILITY INTERVALS FOR BETWEEN-SUBJECTS MEAN PARAMETERS [POPULATION-LEVEL MEANS])

	Page-Request Model (Model C9)	*Page-View-Duration Model (Model D8)*
INTERCEPT	1.885	3.838
	[1.834,1.943]	[3.813,3.867]
BYTES	.211	.558
	[.126,.308]	[.503,.613]
CMAKE	.186*	–.062
	[.105,.269]	[–.102,–.028]
CPAGE	–.711**	.079**
	[–.755,–.668]	[.063,.095]
CSESSION	–.194**	N.S.
	[–.236,–.151]	
DYNAMIC	N.S.	–.400
		[–.450,–.354]
ERROR	.427	.155
	[.317,.542]	[.100,.216]
PREVDUR	–.220*	N.S.
	[–.250,–.189]	
RELOAD	–.050	.373
	[–.102,–.006]	[.344,.404]
SRVRESP	–8.485	8.016
	[–9.434,–7.445]	[5.855,10.272]

N.S. = Did not provide an improvement of log of PSBF; removed from final estimation.
*Enters the model in log form after adding 1 to the variable.
**Enters the model in log form.

variables, we report the percentage change in probability or duration due to the presence of the corresponding factor (e.g., the mean value of –.104 for DYNAMIC in the duration component means that the duration becomes, on average, 10.4% shorter when the requested page contains dynamic content; the mean value of –.020 for RELOAD in the page-request component means that the probability of an additional page request becomes, on average, 2% lower when users reload the page).

From our results, we identified four browsing behavior phenomena for discussion: (1) learning over multiple sessions, (2) within-site lock-in, (3) time-constrained behavior in Web navigation, and (4) the role of cost–benefit trade-offs.

Evidence for Learning Effects

Our findings are consistent with the presence of learning effects that spill over multiple visits. This corroborates Johnson, Bellman, and Lohse's (2003) research, which finds that visitors spend less time per session the more they visit the site. In contrast with their approach, our joint analysis of page requests and page-view durations enables us to explore how this reduction in session duration occurs. We find that the effect of repeat visitation is not the same on the two browsing behavior components. As visitors return to the site, they reduce the number of page views, but there is no significant change in page-view duration. Repeat visits (CSESSION) have a negative impact on the probability of an additional page request. As a result of the log form in which repeat visits enter the binary choice model, this occurs at a decreasing rate, which is consistent with learning effects spilling over into future sessions (i.e., users who revisit a site tend to remember where to go) and the notion that learning need not operate at a linear rate. For example, as users make more visits to the site, they are able to make fewer reductions in their browsing paths. Table 8 gives the cross-sectional means and range of the posterior means of the individual-level elasticities for this variable. The 95% coverage interval for CSESSION is [–.146,–.026], and some visitors have an elasticity as strong as –.3. As noted previously, repeat visits had no significant effect on page-view duration. Thus, the tasks at each page appear to take them as long, even with previous experience.

We attempted to control for alternative explanations by including variables that account for different page content (bytes transferred and dynamic content) and download times (bytes transferred and server response), errors during page transmission, and page reloads, among others. One alternative explanation for fewer page requests on repeat visits is that users perform different tasks when they return, and these tasks require fewer page views. In addition, even if there were learning effects at the level of page-view duration, if users perform somewhat different tasks as they return to the site, we may not be able to capture the learning through the repeat visits variable. To test for this alternative explanation, we included a dummy variable for all page views performed after users placed a car order (ORDER). We expected that tasks performed after ordering are different, and these could lead to a significant change in time spent per page and the probability of requesting additional pages. We found no significant results for this variable in either the choice or the duration components of our model.

We also included variables to control for the user's experience in completing the car configuration task. We tested a

Table 8

INDIVIDUAL-LEVEL ELASTICITIES

	Choice			*Duration*		
	Minimum	*Maximum*	*Mean*	*Minimum*	*Maximum*	*Mean*
BYTES	–.184	.136	.010	.002	.833	.160
CMAKE	–.064	.374	.067	–.688	1.101	–.007
CPAGE	–1.212	–.000	–.238	–.181	.339	.079
CSESSION	–.358	.027	–.067	—	—	—
DYNAMIC*	—	—	—	–.170	.000	–.104
ERROR*	.083	1.578	.836	–.012	.109	.043
PREVDUR	–.524	.077	–.090	—	—	—
RELOAD*	–.124	.091	–.020	.012	.329	.099
SRVRESP	–8.418	.027	–.044	–.373	2.766	.021

*Values denote the percentage change in probability of page request or in duration due to the presence, versus absence, of the factor.

dummy variable set equal to one for page views following a user's first car configuration (FIRSTMAKE) and a variable for the cumulative number of cars configured (CMAKE). First car configuration did not significantly affect either page-request propensity or page-view duration. Cumulative car configurations was significant in both browsing components and was retained in the final models. The parameter means in Table 7 indicate that as users complete the configuration of more cars on the site (i.e., as CMAKE increases), they are more likely to continue browsing but spend less time per page. The sign for CMAKE is opposite that of CSESSION in the page-request model, and the correlation between the two variables is negligible (–.026). This increases our confidence that repeat visits to the site drive its more efficient use.

Although repeat visits do not decrease page-view duration, it is still possible that a learning effect occurs in page-view duration. For example, if returning users are more motivated, they might spend more time per page, which thereby counteracts a learning effect. Without additional information, we cannot disentangle these two effects.

Evidence for Within-Site Lock-In

Our results are also consistent with within-site lock-in. Zauberman (2003) has introduced the term "lock-in" to describe the behavior of subjects he found unwilling to switch to other sites. Zauberman demonstrates that users quickly develop a tendency to become loyal to a single Web site to accomplish an assigned task. This occurs even when better Web sites are just a click away. As with loyalty to a given site, we can envision within-site lock-in. As visitors go deeper into the site on a given visit, they may become more involved. Thus, depth of site visit may lead to greater-than-expected probabilities of page requests and longer page-view durations.

Our results show that visit depth (CPAGE) has a positive impact on mean duration. We can interpret this result as evidence of within-site lock-in for the Web site under analysis; as visitors request more pages, they tend to spend more time viewing each page. Because visit depth enters in log form, this time increase occurs at a decreasing rate. Table 8 gives the mean value of the individual-level posterior elasticities as .079, and the highest positive elasticity is .339. In addition, 26% of visitors have visit-depth elasticity greater than .1, and only 5% of visitors have negative elasticity (less than 2% have negative elasticity greater than .05 in absolute value).

The results from the choice model further support within-site lock-in for this Web site. Visit depth (CPAGE) has a negative impact on additional page requests within the site; however, the log form means that as cumulative page views increase, the negative impact on probability of page request increases at a decreasing rate. Therefore, our results are consistent with a within-site lock-in that does not reverse a stronger effect of time constraints but attenuates it (for the possible effect patterns discussed previously, see Figure 2).

The lock-in result holds even after we control for other factors that could account for this phenomenon, within the limitations of our data set. For example, if the first pages of the Web site (e.g., home page, pages explaining how to use the site) are served by a different server that is faster and more reliable, users could take less time in each of the initial pages simply because the download time and number of errors are lower. We expected the same result if page clutter and number of pictures were lower for the initial pages. In addition, if users who went deeper into the site were also more motivated (e.g., those who want to configure and order cars), we expected a nonlinear result for visit depth in the choice model. Although we cannot control for all possible alternative explanations, in both the choice and the duration models, we used measures extracted from the log files to control (at least partially) for downloading effort, server response time, errors during page download, page reloading by each visitor, tasks performed, and familiarity with configuring cars (BYTES, SRVRESP, ERROR, RELOAD, and CMAKE). In the duration model, we also controlled for the presence of dynamic page content (DYNAMIC), which did not significantly affect page requests. Finally, we accounted for heterogeneity in both models.

Evidence for Time-Constrained Behavior

Time constraints on Internet browsing may be self-imposed or imposed by external environmental conditions. Either way, time constraints likely play an important role in online shopping, as they do in the bricks-and-mortar arena.[4] The strong negative result for visit depth in the page-request model (note the elasticity value for CPAGE in Table 8) shows that as users stay longer at the site, the probability of exiting, given the current page view, increases. Although the within-site lock-in discussed previously attenuates this, the probability of page requests continues to decline as users go deeper into the site. Note that this effect pertains to the probability of site exit conditional on users' arrival at a given page view. This means that even if our model were to predict constant page-request probabilities (which it does not), the probability of site exit would steadily increase for users as they browsed more deeply into the site (analogous to a geometric distribution).

In addition to simply exiting the site, we believe that Internet users may cope with time constraints by more sophisticated mechanisms. For example, visitors might dynamically adjust time spent per page and number of pages requested. Our results suggest that visitors trade off number of pages requested and time spent at each page. Although we found no correlation between the error structures of the two browsing component models (compare the log of PSBF of Model D8, the independent duration model, and the one of Model D9, the full type II tobit model), the results for the feedback effect of duration on page-request choice indicate that the longer a visitor viewed the previous page, the lower the probability was of an additional page request within the site. (From Table 7, the posterior 95% coverage interval for the cross-sectional mean of PREVDUR in the choice model is [–.250,–.189], with a posterior mean of –.220.) This result implies that users may balance the time they spend with each page and the number of page requests so as to avoid exceeding a time budget. We note that few visitors evinced a different behavior. According to the individual-level posterior parameter estimates, only 8.5% of the Web site visi-

[4]Previous research on optimal store location (for a review, see Craig, Ghosh, and McLafferty 1984) and store choice (Bell, Ho, and Tang 1998) shows how shoppers' value of time influences shopping decisions in the bricks-and-mortar environment.

tors are either insensitive to previous page-view duration or respond positively to this variable in terms of page-request probability.

The time constraints of each visitor also could affect the dynamics of browsing behavior. We expected visitors with more stringent time constraints to be more sensitive to visit depth in their page-request decisions; page-request probability for these visitors should decrease more rapidly as they go deeper into the site. These same visitors should be more sensitive to repeat visitation and previous page-view duration while making their page-request decisions (they are more motivated to reduce the total time spent at the site). In contrast, visitors with more stringent time constraints should be less sensitive to visit depth in their page-view-duration decisions; page-view duration should increase less with visit depth for those users with more stringent time constraints.

To determine whether browsing patterns are consistent with this hypothesized pattern, we examined the correlations among the individual-level elasticities for visit depth, repeat visits, and previous page-view duration. These correlations are presented in Table 9 (for all reported correlations, $p < .01$). For the page-request decision, visitors who are more sensitive to visit depth and repeat visits are also more sensitive to previous duration (note that correlations of the corresponding individual-level elasticities are all positive). This indicates that users with more stringent time constraints (as revealed by their sensitivity to PREVDUR and CPAGE) also learn more quickly (as revealed by a greater sensitivity to CSESSION). In addition, visitors who are more sensitive to visit depth in their duration decisions (i.e., visitors who tend to increase page-view durations as they go deeper into the site) are less sensitive to visit depth, repeat visits, and previous duration in their page-request decisions (note the negative correlations in the last line of Table 9).

This pattern of results suggests that browsing behavior is influenced by strategies visitors develop to cope with time constraints. Visitors' dynamic browsing behavior, within and across site visits, seems to be coordinated and working in the same direction (i.e., to save time) regardless of users' specific motivations (simply browsing or real shopping). This indicates that time constraints may play a major role in Internet browsing.

Evidence for a Cost–Benefit View

The results for page-specific covariates (bytes transferred, server response time, and errors) provide support for a cost–benefit view of Web site browsing behavior, similar to the one on which Huberman and colleagues (1998) base their model. For example, increases in the number of bytes transferred (BYTES) increases the probability of an additional page request (information benefits exceeding costs) but also increases page-view duration. Increases in page-view durations then decrease subsequent page-request probabilities (through the negative effect of previous duration in the binary choice model), which serves as a check on the extent to which additional information can lead to more page views and may reflect users' attempts to balance costs and benefits. An increase in server response time (SRVRESP) has a negative effect on page-request probabilities and extends page-view duration. Because time delays raise costs for Web users, the results support the cost–benefit framework. Errors might also be expected to raise costs and lead to a greater probability of site exit. Errors enter the probit model with a positive sign, implying that errors encountered in browsing the site actually increase the likelihood of remaining on the site. However, the error variable is associated with an increase in page-view duration, which in turn decreases subsequent page-request probabilities through the effect of previous page-view duration. Collectively, these findings suggests that Web site visitors weigh the costs and benefits at each page of their visits. Depending on how the expected benefits compare with the expected costs, users make an additional page request or exit the site.

Table 9

CORRELATION MATRIX FOR SELECTED CHOICE AND DURATION MODEL ELASTICITIES

	Choice			*Duration*
	PREVDUR	*CPAGE*	*CSESSION*	*CPAGE*
Choice				
PREVDUR	1.000			
CPAGE	.736	1.000		
CSESSION	.577	.817	1.000	
Duration				
CPAGE	–.198	–.228	–.186	1.000

DISCUSSION

The results from our models of within-site browsing behavior have several implications for Web site managers and designers. In this section, we discuss why managers should be wary of aggregate-level statistics to infer individual-level behavior. We also discuss the implications of our results for Web site design and customization and factors that should be considered when developing metrics to evaluate site performance.

Aggregate-Level Analysis

Managers interested in understanding the dynamics of visitors' browsing behavior within and across site visits might note that summary statistics for some of the variables of interest can be readily computed from log files and site-tracking databases. In particular, it is straightforward to graph the relationship between average page-request probabilities and either visit depth or repeat visits. Similarly, we can graph the relationship between average page-view duration and visit depth or repeat visits. These graphs are displayed in Figure 3.[5] Figure 4 illustrates the modeling results. The figures depict contrasting images of browsing dynamics.

Turning first to the relationship between average page-request probabilities and visit depth, we note that the line in Figure 3, Graph A, is upward sloping (across most of the data range). This suggests that as users browse more extensively on the site, the probability of additional page requests

[5]The y-axis in Figure 3, Graph A, represents the average page-request probability and is computed by considering all page requests that correspond to a second session page, a third session page, and so on. From these, we determined the percentage of pages that were not the last page viewed in the session (i.e., the rate of staying) for each cumulative page value. The y-axis in Figure 3, Graph B, represents the average number of pages requested per session for the first session, second session, and so on, across all users (i.e., the equivalent per-session measure to the average page-request probability, with identical interpretation at the aggregate level).

Figure 3

SUMMARY STATISTICS FROM AGGREGATE-LEVEL SITE USAGE DATA

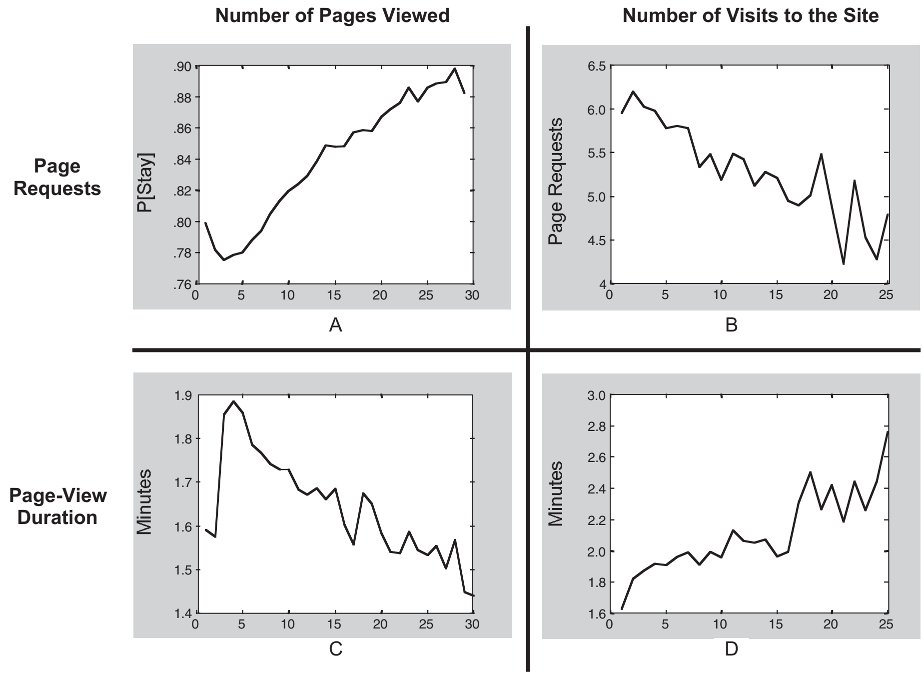
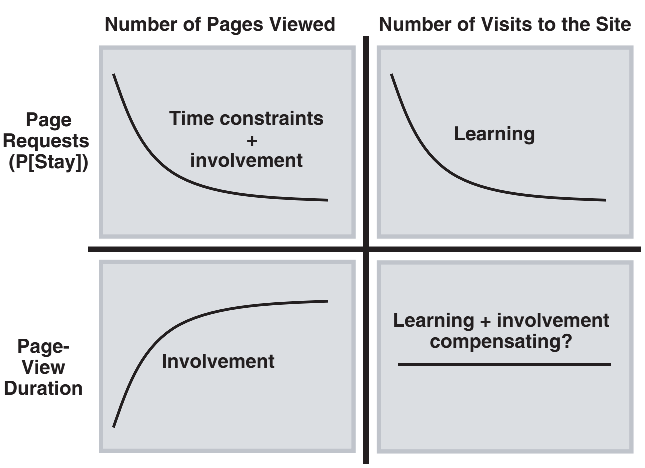


Figure 4

CONCEPTUAL DEPICTION OF THE MODEL RESULTS FOR DYNAMIC BROWSING BEHAVIOR



increases. This indicates a high degree of within-site "stickiness" as more page views are associated with higher page-request probabilities. In contrast, our probit model results show that visit depth is negatively related to page-request probabilities. As a result of the log form of the variable, the negative effects are attenuated as the number of pages viewed becomes larger.

Figure 3 also presents the aggregate-level relationship between average page-view duration and visit depth (Figure 3, Graph C). More page views are associated with shorter viewing times (across most of the data range). Our duration model produces a different result; more page views increase duration (Figure 4). Again, the aggregate-level analysis and the disaggregate-level model produce different implications. However, we note that increases in page-view duration feed back negatively to future page-request probabilities.

The relationship between the average number of pages viewed per session and repeat visits is presented in Figure 3, Graph B, which shows that more site visits are associated with fewer pages requested per session. This result supports a learning hypothesis and is consistent with the modeling results (repeat visits enter the probit model with a negative sign). The log form for repeat visits (CSESSION) suggests that learning effects are subject to diminishing returns. The aggregate-level statistics do not provide this additional insight.

Last, Figure 3 gives the relationship between average page-view duration and repeat visits in terms of the aggregate-level statistics. The line in Figure 3, Graph D slopes upward, indicating that more site visits are associated with longer page-view durations. At first glance, this appears to contradict a learning-based hypothesis (Why are longer page views observed for more experienced users?), but the modeling results show a null direct effect (repeat visits are not significantly related to page-view duration). Putting the two results for repeat visits together suggests that repeat visitors are likely to spend less total time at the site (fewer page requests, no change in duration), which is consistent with a higher proficiency in extracting desired information.

The conflicting results between Figures 3 and 4 suggest that managers should be cautious when using aggregate-level statistics to make inferences about browsing behavior. Several factors can explain these differences. For example, aggregate-level statistics do not account for differences between visitors. Those visitors who make more or deeper visits may browse differently and may respond differently to visit- and page-specific covariates. In addition, aggregate-level statistics provide only a univariate analysis of browsing behavior and do not account for the impact of other relevant factors. Finally, our proposed model can account for the possible interdependence between page requests and page-view-duration decisions that a simple aggregate-level analysis cannot.

Site Design

Our findings also have implications for Web site designers. One implication is that site designers should emphasize reducing the number of page views needed to complete a transaction. This comes from the significant negative effects of visit depth on page-request probabilities. Although visit depth (CPAGE) has a positive impact on duration, duration has a negative feedback effect on future page-request probabilities. Moreover, longer page-view durations do not bring users closer to a completed transaction.

A related implication is the positive impact of bytes transferred (BYTES) on additional page views. This suggests that most users are unlikely to be overloaded with the information presented on the current Web site pages. (Recall that we tested for logarithmic and quadratic effects for this variable in addition to the linear term that appears in the final model.) By taking these two results together, management might consider redesigning pages to place more information on each page but to reduce the number of page views needed to complete a transaction.

Compared with other business-to-consumer e-commerce sites active at the time of our study, the company's Web site was noted for its lack of clutter and its sleek design. In a major redesign of the company's Web site (which occurred approximately three months after we collected our data), various user tasks were moved to earlier page views. The company also introduced a new feature that enables users to compare vehicles side by side on the same page. Although we do not have comparative metrics for the old and redesigned sites, managers reported that they were pleased with the new site's performance.

Finally, our model can help Web site managers perform site changes to accomplish specific goals. For example, managers of Web sites that follow a media model by selling impressions to advertisers (as either their main business model or a parallel one, as in the case of e-commerce sites) may vie for an increase in advertising impressions. One solution has been adopted by many publishing sites: Split tasks over multiple pages (e.g., by splitting articles), and force users to request additional pages to accomplish the same tasks they had been performing. Another alternative is to impose a permanent frame on top of the screen and periodically rotate the banner. The number of page views remains unchanged, but managers can use the viewing time at each page to create multiple exposures.

By fitting our joint model of page requests and page-view duration to the Web site data, we could determine, with a simulation study, which solution to adopt. In the case of the e-commerce site we have analyzed, we can infer from the results that visitors are time constrained and prefer shorter paths to longer ones (e.g., negative impact of visit depth). If site designers were to impose the need for additional page requests, users would respond negatively. For example, this was the actual reaction of Salon.com visitors in May 2000 when confronted with a new site design that promoted more advertising impressions by forcing additional page requests. Eventually, Salon.com lost visitors and traffic, users voiced their discontent throughout Internet chat rooms, and the site returned to the previous design (Eads 2000). Our modeling approach also indicates which other facets of the system to improve or change to increase banner impressions (related to either page requests or viewing duration). For example, our model demonstrates the benefits of decreasing server response time or increasing server reliability (e.g., producing fewer errors) while accounting for the possible interdependence of page requests and page-view-duration decisions.

Site Customization

An additional implication for Web site managers pertains to site customization. Given the heterogeneity in response that we found (see, e.g., the range of the posterior means of the individual-level elasticities in Table 8), a logical next step is to consider how the site might be customized to better serve the idiosyncratic needs of returning visitors. For example, Adar and Huberman (2000) advocate greater use of site customization to exploit differences in Web users' willingness to visit more pages. But our findings suggest that changing the nature of the site structure and organization from one visit to the next (to customize it for a given user) risks attenuating the learning and proficiency effects that may occur, which thereby raises user costs and potentially reduces the number of repeat visits. However, our results do not provide any evidence of whether customizing site content (given a certain basic structure and organization) carries any benefit. To be sure, more attention should be given to the possible drawbacks that arise from site customization. In addition, the advantages of any type of customization need to be carefully measured and weighed against the possible disadvantages (given the purpose of the site).

Metrics

A final implication for managers charged with responsibility for the Web site involves tracking changes in site performance over time. Our findings show that new visitors browse the site differently from repeat visitors (as captured by the repeat visits variable, CSESSION), and extent of repeat visits is also relevant. For example, if the mix of old and new users shifts to new users (e.g., as a result of an advertising campaign that attracts new visitors), the number of page views will increase. This could send a false signal of change in an important Web site metric. Therefore, Web managers should take into account the mix of new versus repeat visitors when analyzing changes in site performance based on aggregate-level statistics.

CONCLUSION

The purpose of this research has been to model the within-site browsing behavior of users of a commercial Web site and the dynamics of this behavior both within and across visits. We focused our study on two elements of browsing behavior: (1) a user's decision to continue browsing the site (by requesting an additional page) or to exit and

(2) how long a user views each page during a site visit. These outcomes of browsing behavior were the dependent variables in our study; they serve to capture some of the key decisions Web users make as they interact with the medium when visiting a specific Web site.

To model these browsing behavior outcomes, we developed individual-level specifications for (1) additional page-request choice (the decision to stay or exit the site) and (2) page-view duration (the decision to view each page for a certain number of seconds). We combine these specifications in a joint model using the type II tobit. The tobit approach naturally handles the structure of the browsing behavior involved (e.g., potential correlation of the outcomes) and the data available from company Web server log files (censoring of the last page-view duration in a given user's session). We estimated the type II tobit model using hierarchical Bayes formulations and MCMC methods, which enabled us to account for cross-sectional heterogeneity in browsing behavior, a phenomenon researchers have already identified as important (e.g., Ansari and Mela 2002; Moe and Fader 2000b), and to investigate the possible relationships between the two browsing behaviors.

We fit our model to server log-file data from a major commercial Web site in the automotive industry. We used a sample of 5000 unique visitors for model estimation. Results show that significant variation in browsing behavior was explained by the covariates in both the page-request and the page-view-duration model components. In particular, our modeling approach enabled us to investigate the dynamics of browsing behavior within and across visits, and we closely examined how visit depth and repeat visits influence the way visitors browse the site. Furthermore, we were able to control for a series of additional factors that may affect browsing decisions.

Substantively, our findings indicate that the Web site users changed their browsing behavior in a manner consistent with learning effects, within-site lock-in, time constraints, and cost–benefit trade-offs. In particular, we found that repeat visits to the site reduce page-request probabilities (resulting in fewer page views) but have no effect on page-view duration. This result corroborates the findings reported by Johnson, Bellman, and Lohse (2003), who show that repeat visitation is associated with shorter total session durations, following the power law of practice. Our results take this further and show that the reduction in session duration stems from fewer page views and not from less time spent viewing each page. We also found evidence in support of the lock-in phenomenon. Unlike previous research on lock-in, our findings pertain to the browsing behavior of users within a particular Web site rather than users' choices made among competing Web sites.

We also showed that inferences drawn from aggregate-level tracking statistics could be misleading. In several cases, the direction of effects was reversed between the aggregate-level relationship and the relationship inferred from the mean parameter of the corresponding covariate in the choice or duration model. Thus, Web site managers should exercise caution when interpreting some of their existing aggregate-level tracking statistics. From a methodological perspective, this highlights the contribution that individual-level modeling approaches may be able to bring to the study of behavior on the Internet. We were able to show how application of the model could provide insights for Web site designers about users' tolerance for more or less information per page and users' willingness to view additional pages. In particular, our results suggest that a relatively clean, uncluttered site improves the site's stickiness properties by placing more information on pages earlier in a user's visit. We note that the trend among electronic retailers to devices such as Amazon.com's "One-Click" ordering is consistent with this finding.

There are several limitations to our work. As an initial modeling study of within-site browsing behavior, there is much that we did not do. We did not incorporate all information about the content of Web pages viewed, and we did not attempt to model users' link choices. We also did not investigate which factors might make a first-time site visitor return to the site and continue returning to the site. Finally, we did not study the users' decisions of whether to place an order through the site but instead kept our focus on browsing behavior per se. We believe the development of new models to address the limitations noted here is a worthwhile area for further research.

APPENDIX A: VARIABLE DESCRIPTIONS

BYTES = sum of the client–server and server–client bytes transferred for a given page (units = 100 kilobytes).

CMAKE = cumulative number of vehicles configured by the visitor up to the current page view.

CPAGE = cumulative number of page views for a visitor to arrive at the current page on a given visit or session.

CSESSION = cumulative number of site visits made by the visitor as of the current page view.

DYNAMIC = dummy variable set equal to 1 if the current page contains dynamic content (e.g., content that requires access to the company's database) and set equal to 0 otherwise.

ERROR = dummy variable set equal to 1 when at least one error occurred during the current page transfer (assuming the visitor eventually succeeds in correctly downloading the page) and set equal to 0 otherwise.

FIRSTMAKE = dummy variable set equal to 1 if the current page was requested at or after the page in which the visitor first configured a car and set equal to 0 otherwise.

ORDER = dummy variable set equal to 1 if the current page view was requested at or after the page at which the visitor ordered a car and set equal to 0 otherwise.

PREVDUR = duration, in minutes, of the immediately preceding page view.

RELOAD = dummy variable set equal to 1 if the visitor reloads the current page at least once (given no errors in transmission) and set equal to 0 otherwise.

SRVRESP = total server response time (10,000ths of seconds).

APPENDIX B: SPECIFICATION OF PRIORS

Model Coefficients

We specify normal priors for the between-subjects conditional mean parameters of the probit and duration components of the type II tobit model (m_{1k}, k = 1, …, p_1, and m_{2r}, r = 1, …, p_2). For the corresponding conditional variances, we specify inverse gamma priors (s^2_k, k = 1, …, p_1, and s^2_r, r = 1, …, p_2). Both the setting of prior hyperparameters and the nature of prior distribution have the potential to influence the posterior distribution of the visitor-specific parameters. In practice, we take diffuse priors to induce a mild amount of shrinkage. All the between-subjects conditional mean parameters share the same prior specification, and similarly we specify identical prior parameters for all between-subjects conditional variances. Therefore, we take

$$\text{(A1)}\quad m_{1k} \sim N(a_0, a_1) \text{ and } s^2_{1k} \sim IG\left(\frac{c_0}{2}, \frac{c_1}{2}\right) \text{ for } k = 1, \ldots, p_1,$$

and

$$m_{2r} \sim N(a_0, a_1) \text{ and } s^2_{2r} \sim IG\left(\frac{c_0}{2}, \frac{c_1}{2}\right) \text{ for } r = 1, \ldots, p_2,$$

where $a_0 = 0$, $a_1 = 50$, $c_0 = 2$, and $c_1 = 2$. These represent diffuse (uninformative) but proper distributions.

Duration Variance and Choice-Duration Correlation

For ρ, the correlation between page-request utility and log of page-view duration, we specify a uniform prior between –1 and 1 (ρ ~ U[–1,1]). For the variance of the log duration, σ^2_2, we specify a diffuse inverted gamma prior such that

$$\text{(A2)}\quad \sigma^2_2 \sim IG\left(\frac{d_0}{2}, \frac{d_1}{2}\right),$$

where $d_0 = 2$ and $d_1 = 2$.

APPENDIX C: FULL CONDITIONALS AND SIMULATION ALGORITHM

Independent Probit and Log-Normal Duration Models

First, set starting values for the unknown parameters.

Second, simulate the utilities (y^*_{1ij} for i = 1, …, N and j = 1, …, J_i), given the parameter values

$$\text{(A3)}\quad y^*_{1ij} \middle| \mathbf{x}_{1ij}, \beta_{1i}, \delta_{ij} \sim \text{truncated } \mathbf{N}\left(\mathbf{x}_{1ij}\beta_{1i}, 1\right),$$

$$\text{such that if } \delta_{ij} = 1 \text{ then } y^*_{1ij} \geq 0, \text{ and}$$

$$\text{if } \delta_{ij} = 0 \text{ then } y^*_{1ij} < 0.$$

Third, draw β_{1ki} for k = 1, …, p_1 and i = 1, …, N from the following conditional posteriors:

$$\text{(A4)}\quad \beta_{1ki} \middle| \beta_{1-ki}, \mathbf{Y}_1, \mathbf{X}_1 \sim \mathbf{N}\left(\overline{m}_{1k}, \overline{s}^2_{1k}\right)$$

$$\text{for } k = 1, \ldots, p_1 \text{ and } i = 1, \ldots, N,$$

$$\overline{s}^2_{1k} = \left(\sum_{j=1}^{J_i} x^2_{1kij} + \frac{1}{s^2_{1k}}\right)^{-1},$$

$$\overline{m}_{1k} = \overline{s}^2_{1k}\left(\sum_{j=1}^{J_i} x_{1kij}\overline{y}_{1kij} + \frac{m_{1k}}{s^2_{1k}}\right), \text{ and}$$

$$\overline{y}_{1kij} = y^*_{1ij} - \sum_{l \neq k} \beta_{1li} x_{1lij}.$$

Fourth, draw m_{1k} for k = 1, …, p_1 from the following conditional posterior distributions:

$$\text{(A5)}\quad m_{1k} \middle| \beta_{1k}, s^2_{1k} \sim \mathbf{N}\left(\overline{a}_{0k}, \overline{a}_{1k}\right) \text{ for } k = 1, \ldots, p_1,$$

$$\overline{a}_{1k} = \left(\frac{N}{s^2_{1k}} + a_1^{-1}\right)^{-1}, \text{ and}$$

$$\overline{a}_{0k} = \overline{a}_{1k}\left(\sum_{i=1}^{N} \frac{\beta_{1ki}}{s^2_{1k}} + \frac{a_0}{a_1}\right).$$

Fifth, draw s^2_{1k} for k = 1, …, p_1 from the following conditional posterior distributions:

$$\text{(A6)}\quad s^2_{1k} \middle| \beta_{1k}, m_{1k} \sim IG\left(\frac{\overline{c}_{0k}}{2}, \frac{\overline{c}_{1k}}{2}\right) \text{ for } k = 1, \ldots, p_1,$$

$$\overline{c}_{1k} = c_1 + N, \text{ and}$$

$$\overline{c}_{0k} = c_0 + \sum_{i=1}^{N}\left(\beta_{1ki} - m_{1k}\right)^2.$$

Sixth, repeat Steps 2–5.

Similar expressions hold for the MCMC algorithm of an independent log-normal duration model. For such a model, set starting values for all model parameters and repeat the equivalent Steps 3–5 of the probit MCMC algorithm, making the necessary adjustments of variables and parameters.

Type II Tobit

First, set starting values for the unknown parameters.

Second, simulate the utilities (y^*_{1ij} for i = 1, …, N and j = 1, …, J_i), given the parameter values

$$\text{(A7)}\quad y^*_{1ij} \middle| \mathbf{x}_{1ij}, \beta_{1i}, \delta_{ij} \sim \text{truncated } \mathbf{N}\left(\mathbf{x}_{1ij}\beta_{1i}, 1\right)$$

$$\text{such that if } \delta_{ij} = 1 \text{ then } y^*_{1ij} \geq 0, \text{ and}$$

$$\text{if } \delta_{ij} = 0 \text{ then } y^*_{1ij} < 0.$$

Third, draw β_{1ki} for k = 1, …, p_1 and i = 1, …, N by sampling from the following full conditionals:

$$\text{(A8)}\quad \beta_{1ki} \middle| \beta_{1-ki}, \beta_{2i}, \mathbf{Y}^*_{1i}, \mathbf{X}_{1i}, \mathbf{Y}_{2i}, \mathbf{X}_{2i}$$

$$\propto \exp\left[-\frac{1}{2s^2_{1k}}\left(\beta_{1ki} - m_{1k}\right)^2\right] \prod_{j=1}^{J_i}\left[1 - \Phi\left(\mathbf{x}_{1ij}\beta_{1i}\right)\right]^{(1-\delta_{ij})}$$

$$\times \left\{\Phi\left[\mathbf{x}_{1ij}\beta_{1i} + \rho\sigma_2^{-1}\left(y_{2ij} - \mathbf{x}_{2ij}\beta_{2i}\right)\left(1 - \rho^2\right)^{-\frac{1}{2}}\right]\right\}^{\delta_{ij}},$$

where $\mathbf{Y}_{1i}$ and $\mathbf{Y}_{2i}$ are ($J_i \times 1$) vectors of simulated utilities and page-view durations, respectively; $\mathbf{X}_{1i}$ and $\mathbf{X}_{2i}$ are ($J_i \times p_1$) and ($J_i \times p_2$) matrices of covariates of the page requests and page-view-duration models, respectively; and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The full conditionals of the individual-level parameters are not of any known form. Several algorithms could be applied to sample from this full conditional distribution. We adopted the ARS algorithm from Gilks and Wild (1992) and Gilks (1992).

Fourth, draw m_{1k} for k = 1, …, p_1 (see Step 4 of the independent probit model MCMC algorithm).

Fifth, draw s^2_{1k} for k = 1, …, p_1 (see Step 5 of the independent probit model MCMC algorithm).

Sixth, draw β_{2ri} for r = 1, …, p_2 and i = 1, …, N. To derive the full conditional of the individual-level duration coefficients, note that we can express $E(y^*_{2ij}|y^*_{1ij} > 0)$ as a simple linear function as follows (see Amemiya 1985, pp. 386–87):

$$\text{(A9)}\qquad y_{2ij} = \mathbf{x}_{2ij}\beta_{2i} + \rho\sigma_2\lambda\left(\mathbf{x}_{1ij}\beta_{1i}\right) + \varepsilon_{2ij}$$
$$\text{for i and j such that } y_{2ij} \neq 0$$

where

$$\varepsilon_{2ij} \sim \mathbf{N}\left(0,\ \sigma^2_{eij}\right),$$

$$\sigma^2_{eij} = \sigma^2_2 - \rho\sigma_2\left[\mathbf{x}_{1ij}\beta_{1i}\lambda\left(\mathbf{x}_{1ij}\beta_{1i}\right) + \lambda\left(\mathbf{x}_{1ij}\beta_{1i}\right)^2\right], \text{ and}$$

$$\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}.$$

Here, $\phi(\cdot)$ is the standard normal density function, and $\Phi(\cdot)$ is the standard normal cumulative density function.

To derive the full conditionals of the individual-level duration model parameters, we perform a variable adjustment as follows:

$$\text{(A10)}\qquad \tilde{y}_{2ij} = \left[y_{2ij} - \rho\sigma_2\lambda\left(\mathbf{x}_{1ij}\beta_{1i}\right)\right]\Big/\sigma_{eij}, \text{ and}$$
$$\tilde{\mathbf{x}}_{2ij} = \mathbf{x}_{2ij}/\sigma_{eij}.$$

We also apply the same full conditionals as in Step 3 of the independent probit model MCMC algorithm, such that

$$\text{(A11)}\qquad \beta_{2ri}\big|\beta_{2-ri}, \beta_{1i}, \mathbf{Y}^*_{2i}, \mathbf{X}_{2i}, \mathbf{Y}_{1i}, \mathbf{X}_{1i} \sim \mathbf{N}\left(\overline{m}_{2r}, \overline{s}^2_{2r}\right)$$
$$\text{for } r = 1, \ldots, p_2 \text{ and } i = 1, \ldots, N,$$

$$\overline{s}^2_{2r} = \left(\sum_{j=1}^{J_i}\tilde{x}^2_{2rij} + \frac{1}{s^2_{2r}}\right)^{-1},$$

$$\overline{m}_{2r} = \overline{s}^2_{2r}\left(\sum_{j=1}^{J_i}\tilde{x}_{2rij}\tilde{\tilde{y}}_{2rij} + \frac{m_{1r}}{s^2_{1r}}\right), \text{ and}$$

$$\tilde{\tilde{y}}_{2rij} = \tilde{y}_{2ij} - \sum_{l \neq r}\beta_{2li}\tilde{x}_{2lij}.$$

Seventh, draw m_{2r} for r = 1, …, p_2 (see Step 4 of the independent probit model MCMC algorithm; similar expressions apply).

Eighth, draw s^2_{2r} for r = 1, …, p_2 (see Step 5 of the independent probit model MCMC algorithm; similar expressions apply).

Ninth, draw ρ by sampling from the following full conditional:

$$\text{(A12)}\qquad \rho\big|\mathbf{Y}_2, \mathbf{X}_2, \mathbf{Y}_1, \mathbf{X}_1, \{\beta_{1i}\}, \{\beta_{2i}\}, \sigma^2_2 \propto \prod_{i=1}^{N}\prod_{j=1}^{J_i}$$
$$\Phi\left[\mathbf{x}_{1ij}\beta_{1i} + \rho\frac{\left(1-\rho^2\right)^{-\frac{1}{2}}}{\sigma_2}\left(y_{2ij} - \mathbf{x}_{2ij}\beta_{2i}\right)\right]^{\delta_{ij}}.$$

This full conditional is not of any known form. To sample from this full conditional, we used the ARS algorithm, as for the individual-level parameters of the probit model (see the preceding Step 3).

Tenth, draw the variance of the log duration, σ^2_2, by sampling from the following full conditional:

$$\text{(A13)}\qquad \sigma^2_2\big|\mathbf{Y}_2, \mathbf{X}_2, \mathbf{Y}_1, \mathbf{X}_1, \{\beta_{1i}\}, \{\beta_{2i}\}, \rho \propto \prod_{i=1}^{N}\prod_{j=1}^{J_i}$$
$$\Phi\left[\mathbf{x}_{1ij}\beta_{1i} + \rho\frac{\left(1-\rho^2\right)^{-\frac{1}{2}}}{\sigma_2}\left(y_{2ij} - \mathbf{x}_{2ij}\beta_{2i}\right)\right]^{\delta_{ij}}$$
$$\times\left[\phi\left(\frac{y_{2ij} - \mathbf{x}_{2ij}\beta_{2i}}{\sigma_2}\right)\Big/\sigma_2\right]^{\delta_{ij}}\left(\sigma^2_2\right)^{-\left(\frac{d_0}{2}+1\right)}\exp\left(-\frac{d_1}{2\sigma^2_2}\right).$$

This full conditional, which is not of any known form, is not log concave with respect to σ^2_2 (a condition for ARS). As an alternative, use Gilks, Best, and Tan's (1995) ARMS algorithm or simply sample for the inverse of the variance.

Eleventh, repeat Steps 2–10.

We monitored the chains for convergence, both graphically and by computing several diagnostic statistics on subsamples of the chain. We also ran several parallel chains to determine whether we obtained significantly different results. In general, we allowed for long chains to run even after convergence. On average, we burned in about 20,000 draws and simulated an additional 8000 draws for analysis. We kept only the fourth draw from each chain. We then used the resulting 2000 draws in our analysis.

APPENDIX D: VARIABLE SELECTION

The PSBF for Model 1 (M_1) with respect to Model 2 (M_2) is given by

$$\text{(A14)}\qquad PSBF_{12} = \frac{\prod_{r=1}^{NT} d_r\left(M_1\right)}{\prod_{r=1}^{NT} d_r\left(M_2\right)},$$

where $d_r(M_i)$ is the cross-validation predictive density of model i for the rth observation and NT is the total number of observations. Gelfand and Dey (1994) show that the MCMC estimate of d_r for a given model M_i can be approximated by

$$\text{(A15)}\qquad \hat{d}_r\left(M_i\right) = B\left[\sum_{s=1}^{B}\frac{1}{f\left(u_r|\Theta^{(s)}, M_i\right)}\right]^{-1},$$

where $f(u_r|\Theta^{(s)})$ is the likelihood contribution of the rth observation given the sth parameter draw of the Gibbs sampler, $\Theta^{(s)}$; B is the number of retained MCMC simulations after a burn-in period and thinning of the chain; and u_r is the dependent variable under analysis. In the case of the choice model, $f(u_r|\Theta^{(s)})$ corresponds to the probit likelihood, and in the case of the duration model, $f(u_r|\Theta^{(s)})$ corresponds to the Weibull density.

We performed all the computations in their logarithmic forms to achieve better numerical accuracy and used the log of the PSBF for model comparison. To determine whether the PSBFs could correctly identify relevant and irrelevant covariates, we also performed synthetic data experiments. The results from these experiments demonstrate the ability of PSBFs to distinguish clearly between significant and nonsignificant covariates (results available on request).

REFERENCES

Adar, Eytan and Bernardo A. Huberman (2000), "The Economics of Surfing," *Quarterly Journal of Electronic Commerce*, 1 (3), 203–214.

Allenby, Greg M. and Peter E. Rossi (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89 (March–April), 57–78.

Amemiya, Takeshi (1985), *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Ansari, Asim, Skander Essegaier, and Rajeev Kohli (2000), "Internet Recommendation Systems," *Journal of Marketing Research*, 37 (August), 363–75.

——— and Carl Mela (2002), "E-Customization," *Journal of Marketing Research*, 40 (May), 131–45.

Bell, David R., Teck-Hua Ho, and Christopher S. Tang (1998), "Determining Where to Shop: Fixed and Variable Costs of Shopping," *Journal of Marketing Research*, 35 (August), 352–69.

Catledge, Lara D. and James E. Pitkow (1995), "Characterizing Browsing Behaviors on the World Wide Web," *Computer Networks and ISDN Systems*, 27 (6), 1065–73.

Craig, Samuel, Avijit Ghosh, and Sara McLafferty (1984), "Models of the Retail Location Process: A Review," *Journal of Retailing*, 60 (Spring), 5–36.

Dellaert, Benedict and Barbara E. Kahn (1999), "How Tolerable Is Delay? Consumers' Evaluations of Internet Web Sites After Waiting," *Journal of Interactive Marketing*, 13 (Winter), 41–54.

DeSarbo, Wayne S. and Jungwhan Choi (1999), "A Latent Structure Double Hurdle Regression Model for Exploring Heterogeneity in Consumer Search Patterns," *Journal of Econometrics*, 89 (1–2), 423–55.

——— and Kamel Jedidi (1995), "The Spatial Representation of Heterogeneous Consideration Sets," *Marketing Science*, 14 (3), 326–42.

———, Youngchan Kim, and Duncan Fong (1999), "A Bayesian Multidimensional Scaling Procedure for the Analysis of Revealed Choice Data," *Journal of Econometrics*, 89 (1–2), 79–108.

———, ———, Michel Wedel, and Duncan K.H. Fong (1998), "A Bayesian Approach to the Spatial Representation of Market Structure from Consumer Choice Data," *European Journal of Operational Research*, 111 (2), 285–305.

Dey, Dipak K., Hong Chang, and Subhash C. Ray (1996), "Bayesian Approach in Model Selection for the Binary Response Data," *Advances in Econometrics*, 11 (Part A), 145–75.

Drèze, Xavier and Fred Zufryden (1998), "Is Internet Advertising Ready for Prime Time?" *Journal of Advertising Research*, 38 (3), 7–18.

Eads, Stefani (2000), "Why Salon's New Look Is More Ad Friendly," *BusinessWeek Online*, (accessed June 2, 2003), [available at http://www.businessweek.com/bwdaily/dnflash/june2000/nf006016.htm].

Gelfand, A.E. (1996), "Model Determination Using Sampling-Based Methods," in *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, eds. London: Chapman and Hall, 145–62.

——— and Dipak K. Dey (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistics Society*, B56 (3), 501–514.

———, S.E. Hills, A. Racine-Poon, and A.F.M. Smith (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association*, 85 (June), 972–85.

——— and A.F.M. Smith (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gilks, W.R. (1992), "Derivative-Free Adaptive Rejection Sampling for Gibbs Sampling," in *Bayesian Statistics 4*, J. Bernardo, J. Berger, A.P. Dawid, and A.F.M. Smith, eds. Oxford: Oxford University Press.

———, N.G. Best, and K.K.C. Tan (1995), "Adaptive Rejection Metropolis Sampling," *Applied Statistics*, 44 (4), 455–72.

——— and P. Wild (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41 (2), 337–48.

Heckman, J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46 (4), 931–59.

——— and B. Singer (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52 (2), 271–320.

Huberman, Bernardo A., Peter Pirolli, James E. Pitkow, and Rajan M. Lukose (1998), "Strong Regularities in World Wide Web Surfing," *Science*, 280 (3), 95–97.

Jain, Dipak C. and Naufel J. Vilcassim (1991), "Investigating Household Purchase Timing Decisions: Conditional Hazard Function Approach," *Marketing Science*, 10 (1), 1–23.

Johnson, Eric J., Steven Bellman, and Jerry Lohse (2003), "Cognitive Lock-In and the Power Law of Practice," *Journal of Marketing*, 67 (April), 62–75.

———, Wendy Moe, Peter Fader, Steven Bellman, and Jerry Lohse (2002), "On the Depth and Dynamics of Online Search Behavior," working paper, Department of Marketing, Columbia Business School, Columbia University.

Kamakura, Wagner A. and Michel Wedel (2001), "Exploratory Tobit Factor Analysis for Multivariate Censored Data," *Multivariate Behavioral Research*, 36 (1), 53–82.

Mandel, Naomi and Eric J. Johnson (2002), "When Web Pages Influence Choice: Effects of Visual Primes on Experts and Novices," *Journal of Consumer Research*, 29 (2), 235–45.

Moe, Wendy W. and Peter S. Fader (2002a), "Capturing Evolving Visit Behavior in Clickstream Data," working paper, Wharton School, University of Pennsylvania.

——— and ——— (2002b), "Dynamic Conversion Behavior at E-Commerce Sites," working paper, Wharton School, University of Pennsylvania.

Nielsen, Jakob (2000), *Designing Web Usability: The Practice of Simplicity*. Indianapolis: New Riders Publishing.

Sahu, Sujit K., Dipak K. Dey, Helen Aslanidou, and Debajyoti Sinha (1997), "A Weibull Regression Model with Gamma Frailties for Multivariate Survival Data," *Lifetime Data Analysis*, 3 (2), 123–37.

Tanner, Martin A. and Wing H. Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82 (June), 528–49.

Wedel, Michel, Wagner Kamakura, Neeraj Arora, Albert Bemmaor, Jeongwen Chiang, Terry Elrod, Rich Johnson, Peter Lenk, Scott Neslin, and Carsten S. Poulsen (1999), "Discrete and Continuous Representations of Unobserved Heterogeneity in Choice Modeling," *Marketing Letters*, 10 (3), 219–32.

Zauberman, Gal (2003), "The Intertemporal Dynamics of Consumer Lock-In," *Journal of Consumer Research*, 30 (December), forthcoming.