# Modeling the Clickstream: Implications for Web-Based Advertising Efforts

Patrali Chatterjee • Donna L. Hoffman • Thomas P. Novak
*Rutgers Business School—Newark & New Brunswick, Rutgers University, Newark, New Jersey 07102-1897*
*Owen Graduate School of Management, Vanderbilt University, Nashville, Tennessee 37203*
*Owen Graduate School of Management, Vanderbilt University, Nashville, Tennessee 37203*
*patrali@newark.rutgers.edu • donna.hoffman@vanderbilt.edu • tom.novak@vanderbilt.edu*

In this paper, we develop an analytical approach to modeling consumer response to banner ad exposures at a sponsored content Web site that reveals significant heterogeneity in (unobservable) click proneness across consumers. The effect of repeated exposures to banner ads is negative and nonlinear, and the differential effect of each successive ad exposure is initially negative, though nonlinear, and levels off at higher levels of passive ad exposures. Further, significant correlations between session and consumer click proneness and banner exposure sensitivity suggest gains from repeated banner exposures when consumers are less click prone. For a particular number of sessions, more clicks are generated from consumers who revisit over a longer period of time, than for those with the same number of sessions in a relatively shorter timeframe. We also find that consumers are equally likely to click on banner ads placed early or late in navigation path and that exposures have a positive cumulative effect in inducing click-through in future sessions. Our results have implications for online advertising response measurement and dynamic ad placement, and may help guide advertising media placement decisions.
(*Advertising and Media Research; Clickstream Data; Computer-Mediated Environments; Online Consumer Behavior; Random-Coefficient Models; Internet; World Wide Web*)

## 1. Introduction

Advertising sponsorship is an important revenue model for firms doing business on the Internet. U.S. Web advertising expenditures are expected to reach $6.3 billion in 2003 (Jupiter Research 2003) and predicted to reach more than $14.8 billion by 2005, even as they represent only a fraction (3.3%) of the total $243 billion in advertising expenditures across traditional vehicles such as TV, print, and direct mail (Jupiter Research 2003). Additionally, the share of Web ad expenditures accounted for by mainstream advertisers is expected to increase from 31% in 2001 to 84% in 2005 (*eMarketer* 2001).

Despite these positive indicators, declining click-through rates, confusion concerning appropriate advertising pricing models, and uncertainty regarding whether traditional advertising metrics are appropriate for new media like the Internet, are contributing to increasing skepticism regarding the value of advertising in this digital medium (Hoffman and Novak 2000). While Web sites still sell advertising using traditional CPM (cost-per-thousand) pricing, advertiser insistence on performance-based pricing that links consumer exposure to advertising with actual market response is forcing the emergence of hybrid pricing models that charge on the basis of click-through in addition to mere exposure (Hoffman and Novak 2000).

The Internet is a unique marketing medium because consumer response to online advertising, typically in


0732-2399/03/2204/0520
1526-548X electronic ISSN

the form of so-called "banner ads," can readily be captured and modeled. When a consumer clicks on a banner ad, a *click-through* is recorded in the server access log. Each time a consumer visits a Web page with an inline ad, a "banner impression" or "ad view" is recorded for the advertising sponsor. As in traditional media, the number of impressions generated depends on exposure to the surrounding editorial content and is thus, in part, under the firm's control. Banner ads accounted for 52% of Web ad revenue in 2000 (*eMarketer* 2001).

However, consumers rarely click on banner ads. Average click-through rates have declined dramatically since the late 1990s; currently, fewer than 3 out of every 1,000 visitors to a Web site clicks on a banner ad (*eMarketer* 2001). Despite industry efforts to improve online advertising effectiveness, plummeting click-through rates remain a concern and are fueling industry speculation that clicks on banner ads are entirely random and cannot be influenced by the marketer (Bicknell 1999). Nevertheless, click-throughs remain an important media pricing metric for the Web.

Even though aggregate click rates are low, rigorous examination of individual consumer click-through behavior is important for several reasons. First, click-throughs are a behavioral and therefore more accountable measure of online advertising, especially compared to mere exposure. Second, even though click rates are low, the absolute number of ad exposures and subsequent click-throughs at high-traffic Web sites are still substantial. Thus, click-throughs can be an important mechanism for driving traffic to advertiser Web sites. Finally, modeling the click-through allows us to address several fundamental issues of both theoretical and practical importance in the nascent area of online advertising response measurement, including the number of times an ad should be displayed, the cumulative and marginal impact on click-through of repeated exposures, and declining click-through rates.

We tackle these issues with an analytical approach to modeling consumer response to banner ad exposures at a sponsored content Web site. Our modeling framework allows us to analyze variation in click probability at each banner ad exposure occasion and account for heterogeneity across consumers and evolution in response across sessions for each individual consumer at the Web site.

Our paper makes two key contributions to the literature. First, we develop, estimate, and test an analytical model of consumer response to online advertising in a dynamic framework. We note that empirical analysis of behavioral outcomes at the microlevel of *each ad exposure occasion* has not been investigated in earlier research for *any* media, simply because the data were not available. The random coefficients logit model with evolution that we specify allows for unobserved heterogeneity and evolution (or change) in click proneness and in responses to banner ad exposures using information that can be obtained from clickstream data. Second, our modeling effort reveals important insights into online consumer behavior in the context of response to Internet advertising that can motivate additional research in this area and impact managerial practice.

The rest of the paper is organized as follows. In §2, we develop a set of testable effects of consumer response to banner ads that follow from theory. We develop our clickstream model in §3 and present model results in §4. In §5, we discuss the implications of our modeling effort for research and practice and conclude with suggestions for future research.

## 2. Theory

We seek to model advertising response in digital environments where consumers navigate through content-laden Web sites with embedded banner advertising. The response variable of interest is whether or not a consumer clicked on a banner ad while navigating the Web site. Such navigations produce a *clickstream* of responses that are highly amenable to modeling. We build a multiperiod model that attempts to capture the differential effects of banner ad exposures on click-through over time, both within a single session and across multiple sessions. The key effects we test in our model are introduced below.

### 2.1. Intrasession Exposure Effects

Prior research on repetition effects in advertising and direct marketing suggest two different patterns of

consumer response to repeated advertising exposures within the same Web session. The first pattern posits that response probability decreases over time. This common effect occurs when consecutive stimuli are independent and the probability of a positive response is assumed to be the same across stimuli (Buchanan and Morrison 1988). The second response pattern holds that initial response probability may be low, but increases with repetition to a maximum level and then diminishes over subsequent repetitions. Berlyne's (1970) two-factor theory provides strong support for this inverted-U relationship between the number of ad exposures and responses. In traditional media, this relationship is caused by two opposing factors. In the initial *wearin* stage, increased response opportunity with each additional ad exposure leads to an increase in affect (Pechmann and Stewart 1989). Subsequently, satiation (or tedium) leads to *wearout*, when each additional ad exposure after wearin has a significant negative effect.

We theorize that wearout dominates in online advertising environments so that for most consumers, there are relatively strong diminishing returns to early repeated exposures that taper off as exposures continue. The rationale for this follows from the fact that the first banner exposure provides sufficient opportunity to elicit a response, similar to print advertising (Calder and Sternthal 1980). The Internet offers consumers relatively more control over the communication and exchange process than has been the case in traditional media environments like broadcast and print. Consumers have both a broader and deeper array of choices about how to receive and interact with communications online (Ariely 2000, Hoffman and Novak 1996, Peterman et al. 1999, Sohn and Leckenby 2001). This direct control extends to control over advertising response. Because consumers largely control their exposure and response to online ads, it follows that those consumers who are most likely to attend and click will do so at the first exposure itself.

Research on visual attention to repeated print ads (Pieters et al. 1999) suggests that consumer control over ad exposure allows them to adapt to advertising repetition by reducing exposure duration. The amount of attention paid to the ad is likely to decline after the first exposure in a monotonically decreasing fashion. This implies that the conditional probability of a click following a string of "failed" banner exposures will decline as the number of banner exposures increases.[1] Industry research supports this idea. Usability studies indicate that once consumers have attended to and recognized a banner ad they learn to ignore it and become progressively insensitive to it (Benway 1998, Schroeder 1998). Additionally, commercial studies of "banner burnout" (*DoubleClick* 1996) show that the probability of click is highest on the first banner ad exposure during a session and decreases thereafter, implying a decline in click probability with each additional banner ad exposure. The *DoubleClick* study, invoked by some online advertisers to reject exposure-based advertising pricing, reveals that beyond four banner ad exposures, the probability of a click is zero.

Although we believe that Web wearout describes the response function for most consumers, under what conditions might we expect wearin to occur? At any given banner exposure occasion, a consumer who failed to click on the ad may have yet to notice it due to "banner blindness" (Benway 1998) or may have noticed the ad but declined to attend to it. Commercial usability analysis (Schroeder 1998) and academic studies (Briggs and Hollis 1997, Dreze and Hussherr 2003) do show that most banner ads go unnoticed, despite the use of attention-grabbing execution features. Further, unlike television ads, banner ads occupy a relatively small portion of the consumer's visual field and can be easily missed even if the consumer is interested in them. This suggests that there may be some consumers for whom additional exposures may actually increase the probability of a more positive response. This is because each additional banner ad exposure increases the probability the ad will be noticed and hence the probability it will ultimately be clicked on. However, these gains in response probability from additional banner

[1] Note that a consumer may notice the banner ad and attend to it but fail to click on it because of time pressure, lack of interest, preoccupation with the content, or their desire to accomplish their original navigation goals (Novak et al. 2000). In those cases, additional banner exposures during the session may or may not immediately increase the click probability.

exposures would ultimately be expected to decrease at higher levels of banner exposure.

The control over banner ad exposure and the presence of wearout means that, for most consumers, we could reasonably expect declining click-through probabilities over repeated banner ad repetitions that would level off after a certain number of exposures. Because repeated exposures to a banner ad are likely to have a negative and nonlinear effect on click probability for a majority of consumers, the *aggregate response to the number of exposures within a given session is likely to be negative and nonlinear*. Because click behavior is primarily driven by immediate relevance, the negative effect due to wearout is expected to dominate over any positive effects due to wearin.

**2.1.1. Banner Location in Navigational Path.** Huberman et al. (1998) model Web browsers as constantly making judgments about the value of clicking on a hyperlink on the current page, based on the value of that page and the uncertainty about the value of the pages not yet seen. They find that, in the aggregate, consumers have a lower threshold for uncertainty at the beginning of the navigation session, when they are more likely to click on hyperlinks that deviate from their navigational path. As browsing depth increases, their threshold for uncertainty increases and consumers are less likely to click on hyperlinks unrelated to navigational goals. *This finding suggests that, other things equal, banner ads displayed earlier in the session will be more likely to be clicked on than those consumers are exposed to later.*

### 2.2. Exposure Effects Across Sessions

**2.2.1. Intersession Time.** The more frequently consumers visit a Web site, the more opportunities exist for the online marketer to expose the consumer to advertising messages and build commitment and loyalty (Hanson 2000). On the Web, intersession time (the length of time between a consumer's visits to the site) is somewhat analogous to store repatronage, but besides measuring repeat visits also captures intervisit duration.

Consumers who revisit after relatively short durations are likely to be more goal-oriented than consumers who revisit after relatively longer intervals (Hoffman and Novak 1996), and arguably, more likely to be familiar with the site organization, content, and advertising. Such consumers are also more likely, then, to be exposed to the same banner ads as in prior session(s) and may be more likely to ignore those ads.[2] We would also expect these goal-directed consumers who return to the site relatively quickly will be more likely to be in the wearout segment discussed earlier (i.e., their click probabilities are decreasing). On the other hand, consumers with longer intersessions times may be more likely to forget ads from prior exposures, making it more likely that they will attend and click on a present visit. These consumers will be more likely to be in the wearin segment (i.e., their click probabilities are increasing). *This argues that longer intersession times on prior visits will lead to a higher click probability in the current session.*

**2.2.2. Prior Session Exposures and Future Session Clicks.** If a consumer was exposed to, but did not attend to, the banner ad in earlier sessions, the preattentive mere exposure effect suggests that at sufficiently high levels, these exposures are sufficient to generate a feeling of familiarity and expectation that may be interpreted as a preference or curiosity for the ad (Janiszewski 1993). This may stimulate attention to the ad in subsequent exposures. Whether this will lead to a click in a future session depends on consumer motivation. Higher levels of curiosity would increase the probability that click will occur.

Some research suggests that if the banner ad is repeatedly noticed (and attended) but not clicked on in earlier sessions, recognition and awareness of the ad stimulus and brand name is generated (Briggs and Hollis 1997). On further exposures of the same ad, if there is no additional information (or motivation) for the consumer to consider, then over-exposure, boredom, and tedium may occur, so clicks in future sessions out of curiosity are less likely to occur. This suggests that ads will benefit more from repetition when consumers do not attend in prior sessions. Because many consumers fail to notice banner ads,

[2] However, as one reviewer pointed out, it is also possible that frequent visitors' relatively greater familiarity with the content means they would be able to devote the cognitive resources attending to the banner.

this effect is expected to be pronounced, particularly across sessions. Within a session, it is reasonable to assume that some consumers will be motivated by immediate relevance. However, this becomes less likely across multiple sessions. In other words, across sessions, we expect the positive wearin effect to dominate over the negative wearout. Note also that the wearout effect occurs under relatively short timeframes, while wearin occurs under longer timeframes. If wearout happens, it happens quickly—typically within a single session. *Thus, we expect that the probability of a click-through in a given session will increase the more banner ad exposures there have been in prior sessions*.[3]

**2.2.3. Time Since Last Click in Prior Sessions.** If a consumer clicks on a banner ad, is there value in exposing her to the banner ad again in future sessions? Prior to the first click, a consumer may be uncertain of the usefulness or entertainment value of clicking on the ad. Once a consumer has clicked, however, some curiosity and uncertainty have been reduced. If click behavior is driven largely by curiosity, then we might expect clicks in prior sessions to be negatively related to clicks in future sessions. However, we assume consumers are motivated to attend to and click on ads and that they are aware that online ad content, especially compared to broadcast and print ads, is dynamic, and extensive, often requiring multiple viewings to fully consume.

Over time, recall of the click experience diminishes. Additionally, there may be a renewed interest in the ad content as sessions pass. Thus, given a click in a prior session, *the longer the time interval since the last click, the more likely there will be a click on the ad in the current session*.

**2.2.4. Repeat Visits.** Individual consumer navigational behavior and click response across sessions evolves over time. Consumers behave more ritualistically initially but become more goal-oriented as they gain more experience online (Novak et al. 2000, Schroeder 1998). This suggests that consumers will be more likely to click on hyperlinks in general during initial visits, becoming less exploratory the more they visit, and hence less click prone as the number of visits to the site increases. This observation is also consistent with Huberman et al.'s (1998) result within a session.

Over time, of course, banner ad sensitivity (i.e., the impact of each additional banner ad exposure during the session on click probability) is expected to decline for all consumers. However, wearout is expected to be fastest for repeat visitors compared to new visitors. That is, *we expect click probability to decline as the number of sessions increases.*

# 3. Modeling the Clickstream

## 3.1. The Data

This study uses consumer clickstream data from a high-traffic sponsored content (or "e-zine") Web site[4] from January 1, 1995 to August 14, 1995.[5] These enhanced clickstream data have ad exposure information (banner ads and clicks) for sponsor ads that were served from the content site server. From January 1, 1995 to August 13, 1995, the Web site required mandatory registration to enter the site. Demographic information was collected the first time a consumer registered at the site and selected a user name/id and a password to be used for future visits to the site. Once registered, the visitor simply logged in for future visits. Demographic information that could be merged with the respondents' clickstream data was not available.

We model ad exposure data for consumers during the mandatory registration period because it is not possible to track and measure consumer exposure to advertising across visits accurately in the voluntary registration period. Only one banner ad was displayed on each page, the best possible situation for an online advertiser. Banner ads were either "hard-coded" or rotated in a predetermined frequency. Because banner delivery was not "smart" or interactive (i.e., served according to the customer's

[3] This assumes a linear form for cumulative banner ad exposures in prior sessions. Treating this variable as a quadratic term added no explanatory power.

[4] Undisclosed at the request of the sponsoring Web site.

[5] See Sen et al. (1998) for an extensive discussion of data available from server access logs.

response history), levels of banner ad exposure were exogenously determined.

Several technical characteristics of clickstream data restrict the modeling scope.[6] First, accurate demographics are difficult to obtain, due to consumer reluctance to provide this information in new online environments (Hoffman et al. 1999). Second, ad execution details for banner ads or active ad pages are rarely collected on an ongoing basis, if at all, owing to the intense challenges facing online businesses operating 24 hours a day, 7 days a week, 365 days a year. Third, because site-centric clickstream data do not contain information on the consumer's activities at external sites (unless the sites are part of an advertising network like DoubleClick), consumer actions after a banner ad click cannot be easily tracked. Finally, banner ad exposures and clicks may not be available for all advertisers at the site because some may be served from the advertiser's Web site server as we discuss below. To make our problem tractable, and because of these data limitations, we leave for future research events that occur subsequent to a click.

**3.1.1. Selection of Sponsors.** There were a total of 3,810 Web pages at the Web site; 3,046 (79.95%) were editorial pages with no ads (or pure editorial pages), 307 (8.06%) were editorial pages with banner ads, and 48 (1.26%) were active ad pages.[7] While this site had 42 advertisers, only 2 advertisers (two high-technology firms, identified as sponsors #15 and #34) ran banners that had no any major executional changes or promotional contests during the study period and had banner and active ad pages on the publisher's Web server. The details of ad placements and exposures generated for these two sponsors are provided in Table 1. Total banner exposures were significantly higher for sponsor #15, despite fewer banner ages. This was most likely due to that sponsor's banner ad placement on entry or gateway pages that had relatively higher traffic than the placement pages for sponsor #34.

[6] Note that some of these issues may be alleviated by more controlled data collection in experimental laboratory environments (e.g., Lynch and Ariely 2000) or in direct marketing contexts with large-scale customer databases (e.g., Chen and Iyer 2002).

[7] The remaining 10.73% of pages (409 pages) were Web site management pages.

**Table 1 Ad Placement and Exposure Details for Sponsors with Fixed Banner Ads**

| Advertiser ⇒ | #15 | #34 |
|---|---|---|
| Number of banner pages | 2 | 6 |
| Banner exposures in mandatory registration period | 1,208,707 | 86,251 |
| Clicks in mandatory registration period | 19,070 | 1,083 |
| Overall share of banner ad exposures at Web Site | 6.87% | 1.41% |
| Overall share of clicks at Web Site | 2.85% | 2.71% |

**3.1.2. Selection of Consumers.** A total of 21,783 unique registered users visited the Web site from January 1, 1995 to August 14, 1995. The daily total of nonunique registered visitors ranged from 14,025 (on 4/26/95) to 42,942 (on 7/27/95), with a daily average of 21,850 (mode = 28,664). See Chatterjee (1998) for further distributional details. A sample of registered consumers was selected according to the following criteria: (1) consumers must have been exposed to banner ads either for advertiser #15 or #34 on more than three sessions during the mandatory registration period (5,326); and (2) visited the site during the calibration and prediction period (3,611). This selection rule yielded 3,611 consumers with 843,565 Web page accesses (34,683 banner exposure occasions) during the seven-month mandatory registration study period. Navigational activity was tracked over 227 days (01.08.95–07.14.95) for model estimation and over 29 days (07.15.95–08.13.95) to test predictive ability.

Preliminary analyses (details in Chatterjee 1998) indicated a highly heterogeneous consumer population, so for example, although most consumers who clicked on a banner ad did so early (within the first three exposures), others did not click following as many as 67 banner exposures. Some consumers clicked on their first session at the site, while others clicked for the first time after as many as 59 sessions. This suggested that the average click rate probability would not be adequate to describe consumer response.

### 3.2. Modeling Click-Through

**3.2.1. The Setup.** We model the probability that consumer $i$'s first click-through during session $s$ will

occur at the $o$th occasion during that session.[8] We assume that the $i$th consumer ($i = 1, \ldots, I$) is exposed to a banner ad at occasions $o = 1, \ldots, O_s$, which occur at the $s$th ($s = 1, \ldots, S_i$) session at the site. For notational convenience we do not include subscripts to indicate consumer and session for occasion, and consumer for session: $o_{is[i]}$ and $s_i$. The number of banner exposure occasions $O_s$ will differ for sessions $S_i$ for each consumer and will also differ across the $I$ consumers. The number of, and spacing between, sessions to the site $S_i$ is expected to differ across the $I$ consumers.

The basis of our model is an unobservable, latent variable $Click^*_{iso}$, which can be interpreted as an index representing consumer $i$'s desire or intention to click when exposed to a banner ad for a sponsor at occasion $o$ during session $s$. In practice, the value of $Click^*_{iso}$ is empirically unobservable; however, we observe its dichotomous realization $Click_{iso}$, the click outcome variable for customer $i$, as follows:

$$Click_{iso} = \begin{cases} 1 & \text{if consumer } i \text{ clicks at occasion } o \text{ during session } s, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The observed dependent variables $Click_{iso} = (Click_{is1}, \ldots, Click_{isos})$ indicate whether consumer $i$ clicked at occasion $o$ ($Click_{iso} = 1$) or not ($Click_{iso} = 0$). We develop our model for a single banner ad on each page. Under the model assumptions, for consumer $i$ during occasion $o$ in session $s$, the click/no click outcomes follow a Bernoulli distribution with parameter $\pi_{iso}$:

$$Click_{iso} \mid \pi_{iso} \sim \text{Bernoulli}(\pi_{iso}), \quad (2)$$

where $\pi_{iso} = \Pr(Click_{iso} = 1)$ is the probability that the $i$th consumer clicks after the $o$th banner exposure in the $s$th session given that the consumer has not clicked in the prior $(o-1)$ banner exposure occasions in the session. The click outcome is observed after the banner ad exposure occurs. Thus, Equation (2) predicts the probability that a click response will occur after a consumer is exposed to a banner ad. $Click_{iso}$ is a function of variables that varies across banner exposure occasions, sessions, and consumers.

Using a logistic parameterization for the hazard rate, we express the probability that consumer $i$ will click on the banner on a given exposure occasion $o$ in the session conditional on not having yet clicked as

$$\begin{aligned} \pi_{iso} &= \Pr[Click_{iso} \mid Click_{is(o-1)}, \ldots, Click_{is1} = 0; X_{iso}, Y_{is}] \\ &= \text{Logit}(a_i + \theta' X_{iso} + \beta' Y_{is} + \lambda' Z_i + \varepsilon_{iso}). \end{aligned} \quad (3)$$

Equation (3) thus refers to the probability of the first click in the session and not to subsequent clicks in the same session and includes the following terms:

(i) *A consumer-specific constant*, $a_i$, the intercept term that affects clicking propensity due to unobserved individual characteristics;

(ii) *Variables varying within and across sessions and across consumers*. $X_{iso}$ represents $K \times 1$ measurements of observed variables that vary over banner exposure occasions within the $s$th session of consumer $i$;[9]

(iii) *Variables varying across sessions and consumers*. $Y_{is}$ is the vector of session-specific variables, i.e., values of variables that vary across sessions for each consumer $i$;[10]

(iv) *Variables varying across consumers*. $Z_i$ is the vector of consumer-specific variables, i.e., those variables that describe consumer $i$ and remain constant across sessions for each consumer.

In the model, $\theta'$, $\beta'$, and $\lambda'$ consist of vectors of coefficients associated with the respective explanatory

[8] Alternatively, we could model the number of banner ad exposures required to generate a click as following a Poisson distribution, with individual Poisson parameters following a gamma distribution across consumers. Because the number of clicks is much fewer compared to no clicks, and the effects of banner exposures during a session and in prior sessions can be difficult to separate, we rejected that approach. Further, it is well documented in the statistical literature that in rare events data, bias in rates calculated using Poisson distribution (versus logistic distribution) can be substantially meaningful with sample sizes in the thousands and in a predictable direction: estimated event rates are too small (King and Zeng 2000).

[9] $Banner_{iso}$, number of banner exposure occasions so far in the session; $Pages_{iso}$, number of pages accessed so far in the session; $Advertiser_{iso}$, dummy variable for advertiser. We pool data for two sponsors. See Appendix A.

[10] $IST_{is}$, intersession duration at session $s$; $TBanner_{is}$, cumulative banner exposures since first-ever visit to the site; $TLClick_{iso}$, time since last click in prior sessions (since we are modeling the first click in a session).

variables. The error term $\varepsilon_{iso}$ is distributed Type I extreme value across consumers, leading to a binary logit model. For more than one banner ad on each page, Equation (3) would be replaced by a multinomial logit formulation. The unit of measurement is banner ad exposure occasion: Each consumer enters the sample when first exposed to a banner ad for the sponsor and remains in the sample until a banner click or a session exit. Consumers who never click-through are included in the likelihood function, but observations that occur after the first click-through are excluded in the likelihood function.

**3.2.2. Consumer Heterogeneity as Click Proneness.** Because online advertising can adapt to respond to an individual consumer's behavior, there is an unprecedented opportunity to segment and target consumers at an individual level. Consumer heterogeneity in banner ad responsiveness arises because of differences in innate tendency to click on ads or "click-proneness" (Briggs and Hollis 1997), as well as from differences in involvement across product categories/brands. In either case, an important issue is whether responsiveness is so heterogeneous that estimates of sensitivities are biased if heterogeneity is ignored.

Different approaches have been discussed in the literature to account for consumer level heterogeneity (Allenby and Rossi 1999). We consider the concept of unobserved heterogeneity, in which individuals may differ in terms of some unmeasured variables that affect the click or no click outcome. Inclusion of individual-specific variables in $Z_i$ can capture part of this variation, but it is almost impossible to identify all of the variables affecting response of an individual at any banner exposure occasion. For instance, information on a consumer's demographics, modem speed, Internet access fees, and online experience could all play a role in affecting ad clicking behavior. Because demographic and related consumer-specific information is not readily accessible from clickstream data, we do not specify any consumer-level covariates in our model. Consequently, in this paper we do not separately estimate $a_i$ (consumer-specific intercept) from $\lambda' Z_i$ (consumer-specific variables). Instead, following Jones and Landwehr (1988), we collect all the consumer-specific influences into a single heterogeneity parameter that we term intrinsic click-proneness, $\alpha_i = a_i + \lambda' Z_i$.

Substituting the variables that can be obtained from the clickstream data under consideration in Equation (3) we have a model capturing intraindividual change in click probability,

$$\begin{aligned}\Pr[Click_{iso} = 1 \mid Click_{is(o-1)}, \ldots, Click_{is1} = 0] \\ = [1 + \exp(\underline{\underline{\alpha_{is}}} + \underline{\underline{\theta^1_{is}}} Banner_{iso} + \underline{\underline{\theta^{1'}_{is}}} Banner^2_{is} \\ + \underline{\underline{\theta^2_{is}}} Pages_{iso} + \underline{\underline{\theta^3_{is}}} Advertiser_{iso} + \underline{\beta^1_i} IST_{is} \\ + \underline{\beta^2_i} TBanner_{is} + \underline{\beta^3_i} TLClick_{iso} + \varepsilon_{iso})]^{-1} \qquad (4)\end{aligned}$$

with variables $Banner_{iso}$, the number of banner exposure occasions so far in the session, $Pages_{iso}$, the number of pages accessed so far, $Advertiser_{iso}$, a dummy variable indicating advertiser, $IST_{is}$, the intersession duration at session $s$, $TBanner_{is}$, the cumulative banner exposures since the first-ever visit to the site, and $TLClick_{iso}$ the time since the last click in prior sessions (see also Footnotes 6 and 7). Variable operationalizations are fully described in Figure 1 and §3.3.1 below.

We note from Equation (4) that the parameters in $\alpha$, $\theta$, and $\beta$ are consumer specific. Given enough observations for each consumer in each session, we could consistently estimate these parameters. However, in practice, there is not an adequate number of observations for each consumer to accomplish this task. We capture different levels of heterogeneity in click response across consumers and evolution of response (or learning behavior) across sessions by allowing the coefficients in $\alpha$ and $\theta$ to evolve across sessions for a given individual and vary across consumers. Because there are multiple banner exposure occasions for the same consumer and in each session, the variance in the unobserved customer-specific parameters induces a nonzero correlation in within-consumer outcomes.

We specify that the parameter coefficients for intrinsic click-proneness ($\alpha_{is}$), response to banner ad exposures ($\theta^1_{is}$, $\theta^{1'}_{is}$), pages browsed ($\theta^2_{is}$), and advertiser ($\theta^3_{is}$) in Equation (4) are drawn from a random distribution, i.e., $\alpha_{is} \sim N(\alpha_{is} \mid \bar{\alpha}, \sigma_\alpha)$ and $\theta_{is} \sim N(\theta_{is} \mid \bar{\theta}, \sigma_\theta)$. The mean gives us the average response of the explanatory variable, while the standard deviation will give us a measure of the heterogeneity in

the response coefficient. We expect the random components of the explanatory variables to covary. This assumption strikes a middle ground between estimating a unique set of parameter coefficients for each consumer and assuming that all consumers are the same, and it is the assumption behind random parameters logit. Furthermore, as the consumer revisits and gains experience with the website, responses to the same variables are expected to change over sessions. The corresponding parameters in Equation 4 are underlined with two lines.

**3.2.3. Deriving the Model.** To capture the evolving nature of the coefficients across sessions for each consumer similar to growth models in the statistics literature (Bryk et al. 1996),

$$\begin{aligned}
\underline{\underline{\alpha_{is}}} &= \alpha^0 + \alpha^1 Session_i + \zeta_{is}^0;\\
\underline{\underline{\theta_{is}^1}} &= \theta^{10} + \theta^{11} Session_i + \zeta_{is}^1;\\
\underline{\underline{\theta_{is}^{1'}}} &= \theta^{1'0} + \theta^{1'1} Session_i + \zeta_{is}^{1'};\\
\underline{\underline{\theta_{is}^2}} &= \theta^{20} + \theta^{21} Session_i + \zeta_{is}^2;\\
\underline{\underline{\theta_{is}^3}} &= \theta^{30} + \theta^{31} Session_i + \zeta_{is}^3.
\end{aligned} \tag{5}$$

In Equation (5), the coefficients $\alpha^1$, $\theta^{11}$, $\theta^{1'1}$, $\theta^{21}$, and $\theta^{31}$ capture the evolution of response parameters across sessions.

Coefficients (in $\beta$) of session-invariant variables in Equation (4) vary randomly across consumers, with density $(f\beta_i \mid \theta)$, where $\theta$ are the true parameters of the distribution. We use a random effects specification to characterize the population in terms of the distribution of coefficients. As above, we expect the random components of session-invariant parameters to covary, independent of session-varying parameters. Further, because there are multiple sessions for the same consumer, the variance in the unobserved customer-specific parameters induces a nonzero correlation in within-consumer outcomes. We specify that parameter coefficients as

$$\underline{\beta_i^1} = \beta^1 + \eta_i^1, \qquad \underline{\beta_i^2} = \beta^2 + \eta_i^2, \qquad \underline{\beta_i^3} = \beta^3 + \eta_i^3. \tag{6}$$

Substituting Equations (5) and (6) into Equation (4), and rearranging, we have

$$\begin{aligned}
&\Pr[Click_{iso} = 1 \mid Click_{is(o-1)}, \ldots, Click_{is1} = 0]\\
&= [1 + \exp\{\alpha^0 + \alpha^1 Session_{is} + (\theta^{10} + \theta^{11} Session_{is})\\
&\quad \cdot Banner_{iso} + (\theta^{1'0} + \theta^{1'1} Session_{is}) Banner_{iso}^2\\
&\quad + (\theta^{20} + \theta^{21} Session_{is}) Pages_{iso} + (\theta^{30} + \theta^{31} Session_{is})\\
&\quad \cdot Advertiser_{iso} + \beta^1 IST_{is} + \beta^2 TBanner_{is}\\
&\quad + \beta^3 TClick_{iso} + \varepsilon_{iso} + \zeta_i^0 + \zeta_i^1 Banner_{iso}\\
&\quad + \zeta_i^{1'} Banner_{iso}^2 + \zeta_i^2 Pages_{iso} + \zeta_i^3 Advertiser_{iso}\\
&\quad + \eta_i^1 IST_{is} + \eta_i^2 TBanner_{is} + \eta_i^3 TLClick_{iso}\}]^{-1}.
\end{aligned} \tag{7}$$

In general, modeling with randomly varying coefficients allows us to separate within-session, within-individual (or across session), and between-individual variation. The $\eta$ and $\zeta$ terms in Equation (7) are not directly observed and enter into the unobserved portion of the utility in the equation, allowing the unobserved portion of the utility to be correlated across occasions for the same consumer. This correlation allows random parameters logit to avoid the IIA problem. The inclusion of error terms in the Equation (5) makes Equation (7) difficult to estimate. If the $\zeta$ and $\eta$ terms were excluded, this becomes a fixed effects model and specification is through interaction terms. Unfortunately, that also implies a deterministic relationship between click probability and the influence of ad exposure variables. The statistical estimator must estimate a model with mixed-level errors, a random specification of coefficients, and a binary dependent variable. Details of model estimation can be found in Appendix B.

**3.2.4. Benchmark Models.** We estimate and contrast this model with three alternative models with simpler heterogeneity structures by specifying that the coefficients in $\alpha$, $\theta$, and $\beta$ remain constant or vary across consumers for a given individual.

*Alternative Model* 1 *(logistic model with nonrandomly varying coefficients across session and no heterogeneity)*: We start with a restricted model in which all heterogeneity has been eliminated by restricting all coefficients in Equation (4), to nonrandomly vary across sessions only, i.e., $\alpha_i = \alpha^0 + \alpha^1 Session_i$, $\theta_i = \theta^0 + \theta^1 Session_i$, and $\beta_i = \beta$ in Equation (4) for all consumers $i$. Hence, we retain all applicable interaction terms with $Session_{is}$ for comparison purposes

but do not allow for heterogeneity across consumers. Estimation reduces to the standard binary logit model using maximum likelihood and is equivalent to assuming that all consumers have similar response coefficients across sessions. Substituting in Equation (7) we have

$$\begin{aligned}\Pr[Click_{iso} = 1 \mid Click_{is(o-1)}, \ldots, Click_{is1} = 0] \\ = [1 + \exp\{\alpha^0 + \alpha^1 Session_{is} + (\theta^{10} + \theta^{11} Session_{is}) \\ \cdot Banner_{iso} + (\theta^{1'0} + \theta^{1'1} Session_{is}) Banner_{iso}^2 \\ + (\theta^{20} + \theta^{21} Session_{is}) Pages_{iso} + (\theta^{30} + \theta^{31} Session_{is}) \\ \cdot Advertiser_{iso} + \beta^1 IST_{is} + \beta^2 TBanner_{is} \\ + \beta^3 TLClick_{iso} + \varepsilon_{iso}\}]^{-1}. \end{aligned} \tag{8}$$

*Alternative Model* 2 *(intercept-specific heterogeneity or linear probability model with random intercepts)*: Here we restrict the heterogeneity of coefficients to the intercept term, intrinsic click-proneness, $\alpha_{is}$ only. Chintagunta et al. (1991) notes that incorporating intercept heterogeneity improves model fit and explanatory power. We let consumers differ in their idiosyncratic click-proneness by specifying the intercept term as the sum of an unobserved component $\alpha^0$, which represents the average click-proneness across all consumers; $\alpha^1$ which represents the change in click-proneness in each session; and a random component $\zeta_{is}^0$, which represents stochastic deviation in an individual's click-proneness in each session relative to the population mean.

$$\underline{\underline{\alpha_{is}}} = \alpha^0 + \alpha^1 Session_{is} + \zeta_{is}^0. \tag{9}$$

The response parameters $\theta_i^k = \theta^k$ remain invariant across consumers. In contrast to the standard logit in Alternative Model 1, the stochastic portion of the model $\zeta_{is}^0 + \varepsilon_{iso}$ is in general correlated across banner exposure outcomes and across sessions for an individual because of the common influence of $\zeta_{is}^0$. Heterogeneity across consumers is captured by the probability distribution of the random variables $\exp(\zeta_{is}^0)$. To maximize the likelihood function we specify a normal distribution for $f(\zeta_{is}^0)$. While this intercept-specific heterogeneity model contains just two sources of uncertainty, the error $\varepsilon$ and random effect $\zeta_{is}^0$ our proposed model has eight represented by terms in $\varepsilon$, $\eta$, and $\zeta$. The random effect indicates that there is additional variation in individual click-proneness beyond that explained by evolution in response across sessions. The equivalent specification for Equation (7) will be

$$\begin{aligned}\Pr[Click_{iso} = 1 \mid Click_{is(o-1)}, \ldots, Click_{is1} = 0] \\ = [1 + \exp\{\alpha^0 + \alpha^1 Session_{is} + (\theta^{10} + \theta^{11} Session_{is}) \\ \cdot Banner_{iso} + (\theta^{1'0} + \theta^{1'1} Session_{is}) Banner_{iso}^2 \\ + (\theta^{20} + \theta^{21} Session_{is}) Pages_{iso} + (\theta^{30} + \theta^{31} Session_{is}) \\ \cdot Advertiser_{iso} + \beta^1 IST_{is} + \beta^2 TBanner_{is} \\ + \beta^3 TLClick_{iso} + \varepsilon_{iso} + \zeta_i^0\}]^{-1}. \end{aligned} \tag{10}$$

*Alternative Model* 3 *(random effects model with consumer-specific heterogeneity but no evolution across sessions)*: We develop a random effects specification without considering across-session evolution. We treat all coefficients in Equation (4) as random over individuals. We assume that coefficients are distributed multivariate normal

$$\begin{aligned}\underline{\alpha_{is}} = \alpha + \zeta_{is}^0, \qquad \underline{\theta_{is}^k} = \theta^k + \zeta_{is}^k, \\ \{\alpha_i, \theta_{is}\} \sim MVN\{(\alpha, \theta_s), (\sigma\Sigma)\}, \end{aligned} \tag{11}$$

with unknown means $\alpha$, $\theta$, and variance-covariance matrix $\sigma^2\Sigma$, where the error variance $\sigma^2$ is constant across individuals. We allow the variance of the distributions for each intercept and parameter of the explanatory variables to be different. Thus, the parameters for consumer $i$ vary from the mean through the random additive components $\eta_i$ and $\zeta_{is}^k$. This model is a special case of our proposed model in that, for a given consumer, the parameters remain constant across sessions. Note that the mean of the random components other than zero is not identified, and we expect the random components of the explanatory variables to covary. The random effects model corresponding to Alternative 3 is specified by setting terms with variable $Session_{is}$ in Equation (7) to 0:

$$\begin{aligned}\Pr[Click_{iso} = 1 \mid Click_{is(o-1)}, \ldots, Click_{is1} = 0] \\ = [1 + \exp\{\alpha^0 + \theta^{10} Banner_{iso} + \theta^{1'0} Banner_{iso}^2 \\ + \theta^{20} Pages_{iso} + \theta^{30} Advertiser_{iso} + \beta^1 IST_{is}\end{aligned}$$

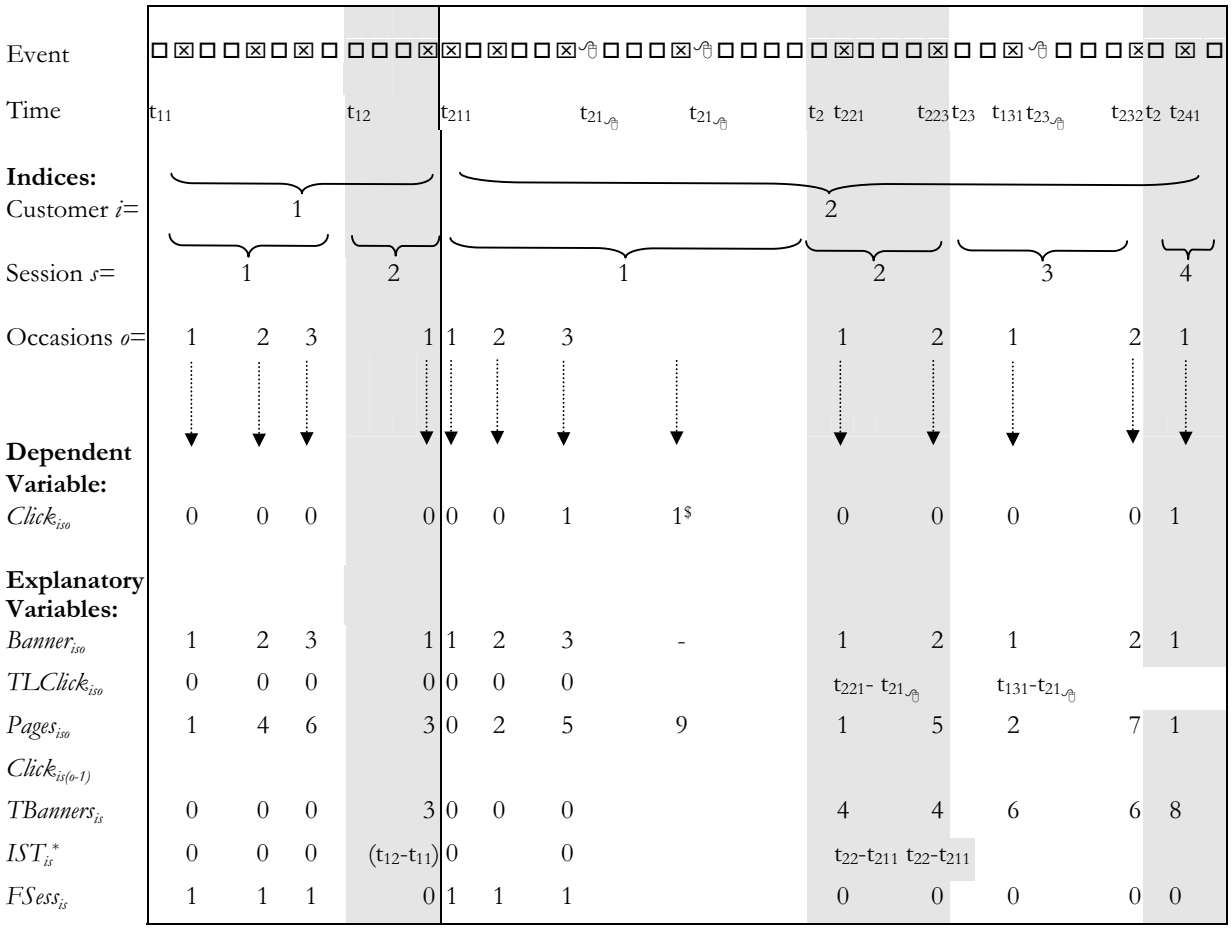
$$+\beta^2 TBanner_{is}+\beta^3 TLClick_{iso}+\varepsilon_{iso}+\zeta_i^0$$
$$+\zeta_i^1 Banner_{iso}+\zeta_i^{1'} Banner_{iso}^2+\zeta_i^2 Pages_{iso}$$
$$+\zeta_i^3 Advertiser_{iso}+\eta_i^1 IST_{is}+\eta_i^2 TBanner_{is}$$
$$+\eta_i^3 TLClick_{iso}\}]^{-1}. \quad (12)$$

### 3.3. Model Specification

The key time-dependent events for our individual-occasion-level model are diagrammed schematically in Figure 1. The figure shows consumer behavior at the website, in terms of page hits (☐ page with no ad, ☒ page with ad) and banner clicks (🖱). Consumers are separated by vertical lines and sessions for each consumer are shaded. Figure 1 shows clearly the complexities involved in constructing variables for clickstream modeling.

**3.3.1. Variable Measurement.** We construct the explanatory variables as follows. $Banner_{iso}$ is the number of times consumer $i$ has been exposed to an advertiser's banner ad in session $s$ so far (i.e., until occasion $o$). We include a quadratic term of banner ad exposures $(Banner_{iso})^2$ to test for the curvilinear effect of repeated banner ad exposures. The number of pages already browsed during the session $Pages_{iso}$ is the total number of pages browsed at the site (including pages that did not have banner ads) during the session till $o$. We capture any systematic differ-

**Figure 1 Schematic Diagram of the Key-Time-Dependent Events for Clickstream Modeling**




*Note.* ☐: Editorial page with no sponsor ad; ☒: page with banner ad; 🖱: banner ad was clicked; $: second click in the session excluded from our model.

ences in click probabilities across the two advertisers or brands by a dummy variable $Advertiser_{iso}$ (=1 for advertiser #15, 0 otherwise).

The session-specific variables corresponding to time since last click in prior sessions ($TLClick_{iso}$), intersession time ($IST_{is,}$), number of times the consumer has visited the site ($Session_{is}$) and cumulative banner exposures in prior sessions ($TBanner_{is}$) are measured for each session and remain the same for all banner exposure occasions in a session. The time since last click in prior sessions $TLClick_{iso}$ is the logarithm of time since last click in previous sessions at the site. If the consumer clicked more than once in a session, then the last click could potentially be in the current session. Because we are modeling first click in a session, $TLClick_{iso}$ is always from the prior session and greater than 0 if the consumer has ever clicked, by definition. Figure 1 shows that if the consumer never clicked on an ad, $TLClick_{iso}$ is set to zero. This does not represent the true time since last click, but it does serve to eliminate this term from Equation (7). $TLClick_{iso}$ is the time since the last of the multiple clicks in the prior session. For example, $TLClick_{iso}$ for the first banner ad exposure occasion in session 2 for consumer 2 is $t_{221} - t_{21}$.

The number of times a consumer visited the site, including those visits where there was no exposure to banners for the advertiser ($Session_{is}$) and the cumulative number of banner exposures in prior sessions ($TBanner_{is}$) is measured since the time the consumer first registered at the site till the occasion under consideration from registration records. These can be calculated a priori from consumer registration records and clickstream history at the site.

# 4. Results

Descriptive statistics for the explanatory variables are in Table 2. Because click occurrences are rare events, pooling the data offers distinct advantages over modeling each sponsor separately. We tested for overall homogeneity of both sponsors and concluded that pooling was appropriate. The details are provided in Appendix A.

We first tested to see if the data supported specifying the explanatory variables as random. Preliminary

**Table 2 Descriptive Statistics**

| Variables | Sponsor #15 | Sponsor #34 | Pooled Data |
|---|---|---|---|
| *Occasion-Varying* | | | |
| Average number of banner exposures in session $s$ until $o$: $Banner_{iso}$ | 4.93 (7.63) [1–90] | 3.53 (1.19) [1–15] | 3.66 (6.9) [1–90] |
| Pages browsed so far in session: $Pages_{iso}$ | 3.65 (4.54) [1–132] | 2.17 (5.83) [1–84] | 3.22 (5.01) [1–132] |
| Time since last click (logarithm hour): $TLClick_{iso}$ | 2.32 (1.81) [0–5.63] | 0.75 (1.46) [0–0.69] | 1.90 (1.86) [0–5.63] |
| *Session-Varying* | | | |
| Average intersession time (minutes) $AIST_{is}$ | 85.11 (96.66) [3.03–1,615] | 88.03 (108.24) [3.15–1,454] | 85.69 (99.07) [3.03–1,615] |
| Number of cumulative banner exposures in prior sessions: $TBanner_{is}$ | 20.05 (33.4) [0–281] | 17.43 (30.84) [0–173] | 19.54 (33.01) [0–281] |
| Percentage of clicks in first session | 15.51 | 3.48 | 11.17 |
| *Other Information* | | | |
| Number of clicks | 1,107 | 624 | 1,731 |
| Number of banner exposure occasions | 23,974 | 10,709 | 34,683 |
| Number of sessions with banner exposures | 4,704 | 3,629 | 8,333 |

*Note.* Standard deviation is in parentheses ( ); range is in square brackets [ ].

estimation of reliabilities indicated that the random effects of quadratic term of banner exposure $(Banner_{iso})^2$ and advertiser $(Advertiser_{iso})$ across consumers are not significant; hence we treat them as fixed effects. Similarly, analyses of the reliability variance estimates suggests that random effects of session-level variables time since last click in prior sessions $(TLClick_{iso})$ and total banner exposure $(TBanner_{is})$ in Equation (7) should be constrained to zero, hence these variables are also specified as fixed effects. Hence Equation (4) can now be respecified as

$$\begin{aligned}\Pr[Click_{iso} = 1 \mid Click_{is(o-1)}, \ldots, Click_{is1} = 0] \\ = [1 + \exp(\underline{\underline{\alpha_{is}}} + \underline{\underline{\theta^1_{is}}} Banner_{iso} + \theta^{1'}_{is} Banner^2_{is} \\ + \underline{\underline{\theta^2_{is}}} Pages_{iso} + \theta^3_{is} Advertiser_{iso} + \underline{\beta^1_i} IST_{is} \\ + \beta^2_i TBanner_{is} + \beta^3_i TLClick_{iso} + \varepsilon_{iso})]^{-1}. \quad (13)\end{aligned}$$

### 4.1. Model Fit

Table 3 reports the fit statistics for the four models. We used the log likelihood, defined as $\bar{\rho}^2 = 1 - (L - k)/L(0)$, where $L$ is the log-likelihood of the model being estimated, and $L(0)$ is the log-likelihood of the model with only the intercept term, thus adjusting for the number of parameters in each model (Ben-Akiva and Lerman 1985) to compare predictive performance of the proposed models. By adding heterogeneity and evolution in the intercept, the log-likelihood increases by over 20% in Alternative Model 1. The advantages of accounting for heterogeneity only in the intercept and slope parameters in Alternative Model 2 and heterogeneity and evolution in the slope and intercept coefficients in our proposed model are highlighted by the significant improvement in fit.

We also report the Akaike Information Criterion (AIC $= -LL + k$). The advantage of BIC over AIC is that it penalizes for an increase in the number of parameters and sample size. After accounting for the increased number of parameters via the BIC (Bayesian Information Criterion; defined as $-LL + 0.5k \log(N)$, where $k$ is the number of parameters, $N$ is sample size, and $LL$ is the log-likelihood), our proposed model incorporating heterogeneity and evolution across sessions in the intercept and slope parameters is the preferred specification.

Because the behavior we are trying to predict is relatively rare (the base probability of outcome is very low), we also calculated another measure of predictive fit as described in Morrison (1969). We rank ordered the 11,619 observations in the holdout sample in decreasing order of their predicted probabilities and classified the first 561 as clicks (the total number of clicks observed in the holdout sample). We also

**Table 3 Prediction Success Table for Click Outcomes**

| | Predicted Outcomes | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Proposed Model | | Alternative 1 | | Alternative 2 | | Alternative 3 | |
| Observed Choice | Click | No Click | Click | No Click | Click | No Click | Click | No Click |
| Click (561) | 232 | 329 | 14 | 517 | 136 | 425 | 187 | 374 |
| No click (11,058) | 64 | 10,994 | 221 | 10,837 | 272 | 10,785 | 139 | 10,919 |
| Total (11,619) | 316 | 11,303 | 235 | 11,354 | 408 | 11,210 | 296 | 11,357 |
| Hit rate | 41 | 99.4 | 2.4 | 95.4 | 24.2 | 96.2 | 33.3 | 98.7 |
| Total hit rate | 96 | | 93 | | 94 | | 95.5 | |
| Success index | 0.279 | | 0.170 | | 0.213 | | 0.256 | |
| Log-likelihood | −13,245.7 | | −14,247.8 | | −13,801.4 | | −13,598.7 | |
| Fit Statistics for Calibration Sample | | | | | | | | |
| Log likelihood | −39,166.4 | | −43,678.9 | | −41,139.7 | | −40,016.5 | |
| AIC | 39,184.4 | | 43,689.9 | | 41,151.7 | | 40,030.5 | |
| BIC | 39,207.45 | | 43,703.99 | | 41,167.07 | | 40,048.43 | |
| $\bar{\rho}^2$ | 0.36 | | 0.28 | | 0.32 | | 0.34 | |
| No. of parameters estimated | 18 | | 11 | | 12 | | 14 | |

report success indices for each of the models. While the alternative models also have high hit rates, this prediction accuracy should be interpreted with caution (because it is driven by disproportionately higher numbers of no click outcomes), given that only 2.4%, 24.2%, and 28% of clicks are correctly predicted. The proposed model actually does far better in predicting clicks (41%), the gain being primarily due to incorporating correlated random effects and evolution across sessions.

Note that from a managerial standpoint, a simple linear additive model will correctly predict the average click rate, though this predictive accuracy will come at the expense of diagnostic ability.[11] Because click-throughs on banner ads are extremely rare events, with a large number of nonevents (i.e., no clicks) and very few events (i.e., clicks), it is well known that logistic regression models can sharply underestimate the probability of occurrence of events (e.g., see King and Zeng 2000).

Table 4 reports the analyses of consumers' click decisions using Equations (7), (8), (10), and (12) modified according to respecifications in Equation (13). Parameter estimates and standard deviations (in parentheses) are reported for variables significant in at least one model. The parameter estimates of the proposed model are significant and in the expected direction. Similar patterns obtain for the alternate models, though differing parameter magnitudes would lead to different managerial conclusions.

### 4.2. Model Implications

In this section, we discuss our specific findings. Within a session, we expected a negative and nonlinear effect on click probability due to wearout; we also theorized that earlier ads would have a higher probability of being clicked on than ads exposed later.

Across sessions, we expected that longer intersession times in prior sessions and more banner exposures in prior sessions, and more time since the last click in prior sessions would lead to higher click probabilities in the current session. Additionally, we expected that click probability would decline as the total number of sessions increased. We also discuss our results for consumer heterogeneity in terms of click-proneness.

**4.2.1. Intrasession Effects.** Both the variable banner ad exposures $Banner_{iso}$ and its quadratic term $(Banner_{iso})^2$ have a significant effect on probability of first click in a session in all of the model specifications. However, Table 4 shows that the coefficient of linear effect of banner exposures is negative and significantly larger than the positive quadratic term leading to a negative and nonlinear impact on click probability, as expected. Most consumers click on the first exposure to the banner ad in a session. If a consumer does not click on the first banner ad exposure, additional banner exposures in the session have lower probabilities of generating clicks initially, but this negative effect levels off at very high levels of banner ad exposure. This indicates that the marginal effect of banner ad exposures on click probability is negative at an increasing rate until the tenth banner exposure in the session and decreasing thereafter.

In aggregate, the elasticity function reaches its minimum at 11 exposures, increasing thereafter indicating there might be incremental gain in displaying banner ads for the sponsor more than 11 times during a session. In our study, consumers were exposed to more than 11 banner ad exposures in 21.5% of all sessions. Each additional banner ad exposure decreases the click probability by a factor of 0.672 (on a base probability of 0.043) until 11 exposures and increases thereafter, within the range of our empirical data. Note that banner exposure coefficient is a random effect that changes from session to session for each consumer (however, the quadratic coefficient does not, the data support its specification as a fixed effect). The statistical significance of the variance of the banner exposure coefficient ($\mathrm{Var}(\theta^{10}) = 1.082$) indicates that consumers are, as we theorized, heterogeneous in their response to banner ad exposures.

The mean of number of pages browsed in the session $s$ till banner exposure occasion $o$ was not significantly different from zero at 90% confidence ($\theta^{20} = -0.013$); however, the standard deviation of the coefficient ($\mathrm{Var}(\theta^{20}) = 1.094$) is significant and fairly large. This suggests that number of pages browsed affects the click decision, with some consumers preferring to click early during the session and others clicking late

[11] We thank the AE for suggesting this.

**Table 4 Probability of First Click in a Session: Regression Coefficients**

| Models: | Proposed Model | Alternative 1 | Alternative 2 | Alternative 3 |
|---|---|---|---|---|
| Evolution across sessions: | Yes | None | Intercept only | None |
| Heterogeneity specification: | Intercept and slope | None | Intercept only | Intercept and slope |
| Variables | | | | |
| $\alpha^0$ | **−4.123** | **−4.016** | **−4.623** | **−4.003** |
| Click-proneness intercept | (**0.379**) | (0.381) | (0.470) | (0.454) |
| $Banner_{iso}$ | **−0.402** | **−0.378** | **−0.266** | **−0.307** |
| Banner ad exposures present session | (**0.071**) | (0.074) | (0.069) | (0.104) |
| $(Banner_{iso})^2$ | **0.018** | **0.016** | **0.014** | 0.0006 |
| Quadratic effect of present session banner exposures | (0.009) | (0.001) | (0.004) | (0.004) |
| $Pages_{iso}$ | −0.013 | **−0.040** | −0.007 | −0.011 |
| Number of pages browsed | (**0.086**) | (0.016) | (0.016) | (0.013) |
| $IST_{is}$ | **0.131** | **0.113** | **0.276** | 0.035 |
| Intersession time till session $s$ | (**0.073**) | (0.033) | (0.040) | (0.037) |
| $TBanner_{l(s-1)}$ | **0.052** | **0.033** | **0.024** | **0.032** |
| Cumulative banner ad exposures in prior sessions | (0.002) | (0.002) | (0.003) | (**0.002**) |
| $TLClick_{iso}$ | **0.391** | **0.377** | **0.318** | **0.316** |
| Time since last click | (0.039) | (0.021) | (0.021) | (0.021) |
| $Session_{is}$ | **−0.006** | **0.020** | **0.018** | — |
| Number of sessions at site | (0.002) | (0.007) | (0.005) | |
| $Banner_{iso} * Session_{is}$ | **−0.021** | **−0.014** | **−0.013** | — |
| | (0.009) | (0.006) | (0.005) | |
| $Pages_{iso} * Session_{is}$ | **0.014** | **−0.112** | **−0.114** | — |
| | (0.016) | (0.019) | (0.017) | |
| $Advertiser_{iso}$ | 0.002 | 0.019 | **0.045** | 0.003 |
| | (0.005) | (0.119) | (0.097) | (0.106) |
| Variance of random effect intercepts | | | | |
| $\alpha^0$ | **7.37** | — | **6.94** | **3.87** |
| | (0.609) | | (0.283) | (0.922) |
| $\theta^{10}$ | **1.082** | — | — | **2.965** |
| | (0.180) | | | (0.313) |
| $\theta^{20}$ | **1.094** | — | — | **0.149** |
| | (0.011) | | | (0.015) |
| $\mathrm{Cov}(\alpha^0, \theta^{10})$ | **−2.107** | — | — | **−4.99** |
| | (0.342) | | | (0.547) |
| $\mathrm{Cov}(\theta^{10}, \theta^{20})$ | 0.017 | — | — | **0.111** |
| | (0.034) | | | (0.049) |
| $\mathrm{Cov}(\alpha^0, \theta^{20})$ | **−0.311** | — | — | **−0.062** |
| | (0.074) | | | (0.100) |
| $\beta^1$ | **2.014** | | | |
| | (0.037) | | | |

*Notes.* Data was pooled for both sponsors. The log-likelihood value for the model with only intercepts was −60,835.91. The normalization in the estimation was with respect to advertiser #34. **Boldface** type indicates that probability exceeds 0.90 that coefficient is $>$ or $<0$, as indicated by the sign of coefficient.

in the session. The mean is not significantly different from zero because the different behaviors tend to cancel each other out in the population.

**4.2.2. Intersession Effects.** The positive coefficient of intersession time ($\beta^1 = 0.131$) on click probability indicates that for each consumer, click probability increases with increasing duration between visits, as expected. The large, statistically significant variance ($\text{Var}(\beta^1) = 2.014$) suggests time between visits significantly impacts click behavior on ads and that consumers differ with respect to repeat site visit behavior. In general, new visitors and less frequent visitors are more likely to click on ads than more regular visitors. Between-session variation is large, most likely because consumers differ in their goals and orientation for visiting a site and contextual factors such as time pressure, also differ on each visit.

Also as expected, the effect of cumulative banner exposures in prior sessions has a small positive, but significant, effect on click probability. Each additional banner ad exposure increases the click probability in future sessions by a factor of 1.003.

Consumers who clicked on a banner ad at least once in prior sessions had a significantly higher propensity to click after exposure to the banner ad, compared to those who never clicked ($\beta^2_{is} = 0.391$, Table 4). Additionally, consumers who have already clicked in earlier sessions are more likely to click in future sessions as time since last click increases, as expected. For each additional day since the last click, the predicted click probability increases by a factor of 1.497 (on a base probability of 0.041, ignoring the effect of other variables).

We find that, overall, click probability decreases with increasing visits to the site (Table 4, $\alpha^1 = -0.006$). This negative repeat visit effect was expected due to increases in experience and consumer learning. However, split-sample results (not reported here in the interests of space) indicate this may not always be true. Click-through rates may increase with familiarity.[12]

**4.2.3. Click-Proneness.** The click-proneness intercept $\alpha^0 = -4.123$ (Table 4, second row) suggests a click probability of 0.039, after all other explanatory variables are set to zero, i.e., when click probabilities depend solely on intrinsic characteristics of consumers. As the click-proneness intercept becomes increasingly negative, the consumer click probability decreases. The large variance of the click-proneness intercept indicates significant dispersion in click-proneness across consumers.

The correlations among variables specified as random effects across consumers (see Table 4) has important implications. Across consumers, the estimated correlation between banner exposure coefficient and the click-proneness coefficient ($\text{Cor}(\alpha^0, \theta^{10}) = -0.74$) is significant, providing evidence that response to each additional banner ad exposure varies with innate click-proneness of the consumer. Banner ads wear out faster for consumers with a higher click-proneness coefficient. The small but significant negative correlation between click-proneness and number of pages browsed so far ($\text{Cor}(\alpha^0, \theta^{20}) = -0.109$) indicates that consumer with higher click-proneness coefficient browse through fewer pages than those with lower click-proneness coefficient.

## 5. Discussion

In this paper, we modeled the clickstream of consumer responses to banner ads at an advertiser-supported Web site with mandatory visitor registration. Our parsimonious, yet flexible modeling approach allows us to decompose the variability in consumers' binary click responses over time and demonstrate that the conditional probability of a click response is heterogeneous across consumers and varies across sessions for each consumer. Within a session, we found a negative and nonlinear effect on click probability due to wearout, and that earlier ads had a higher probability of being clicked on than ads exposed later. Across sessions, we found that longer intersession times in prior sessions, more banner exposures in prior sessions, and more time since the last click in prior sessions led to higher click probabilities in the current session. We also found that click probability declined as the total number of sessions increased.

[12] We thank the AE for pointing this out.

### 5.1. Consumer Identification and the "Cookie" Problem

The lack of complete consumer identification in clickstream data poses a challenge that many advertiser-supported Web sites address by implementing identification procedures involving "cookies" (small files stored on the consumer's hard disk that allow the client-side browser to be tagged with a unique identifier) or voluntary registration. While we use clickstream data collected at a Web site with a mandatory registration policy over a sufficiently long period of time, most Web sites are reluctant to require mandatory registration. However, because cookies operate implicitly, most consumers are not aware[13] that virtually all commercial Web sites use cookies to track visitor behavior.

As a modeling solution, the use of cookies to track consumer behavior across sessions is problematic. Many consumers access the Web using multiple browsers on multiple computers (e.g., home, school, office, friend's house, library, and so on). Additionally, a single computer (e.g., at home) may have multiple users, and each time a computer accesses the Internet through an Internet service provider, dynamic addressing assures that each session is assigned a different IP address. Thus, using only cookies, it becomes very difficult to assign a particular consumer to a particular session, let alone link sessions together. One popular technique captures the consumer's name at some point through registration or purchase and subsequently uses this information to identify the individual and attempt to link sessions, for example by displaying something like, "Welcome back, Jane Doe. If you are not Jane Doe, then please log in here." Yet clearly, cookies do not allow the modeler to exploit similarities in click behavior across sessions for each consumer and further restrict heterogeneity at the level of each consumer session. This therefore has the potential to create biased estimates of response parameters.

Our research highlights the importance of explicit consumer identification procedures to enhance the value of clickstream data. Intersession click behavior was an important component of our model across all segments for ads of both sponsors in this study and employed to capture heterogeneity across consumers and model click propensity. Without such identification, the modeling effort will necessarily be limited to cross-sectional effects and may lead to considerably weaker performance in terms of predictive ability.

Although clickstream data undoubtedly represent a powerful new source of behavioral insight for marketing scientists, the clickstream data we analyze in this paper limit our modeling effort in a number of ways. First, we must be cautious generalizing from just two sponsors. Second, our data preclude determination of whether previous clicks on *any* ads in prior sessions, compared to the identical or even similar ads in prior sessions, increases the click rate in the current session. Third, the availability of banner ad exposures and click data on all advertisements at the site would have permitted estimation of overall click-proneness for each consumer at the site, thus allowing the site to identify consumers who click more on ads in general. Fourth, data on individual banner executions, ad/content congruence and content refreshment rates would possibly yield richer insights regarding observed wearout effects. Finally, post-click response data, not available for this study, would also have been useful. Nevertheless, we have obtained a number of interesting theoretical results with important implications for online advertising practice.

### 5.2. Key Managerial Implications

In stark contrast with broadcast media, online marketers are able to use sophisticated technologies such as smart ad delivery and tracking on rich clickstream databases to fit models similar to the ones developed here. Such individual-level models can thus help marketers develop more effective online advertising strategies.

Our results tend to show that the more click-prone consumers experience faster wearout to banner ads and browse through fewer pages than less

[13] Experienced online consumers are aware that they can change the default settings of cookie preferences on their browsers to disable cookies if they so desire. In that case, using cookies becomes explicit. However, most advertiser-supported commercial Web sites deny access to browsers that have disabled cookies, so in practice, the consumer has no choice but to allow cookies if Web site entry is desired.

click-prone consumers. The marketing implications are simple but important. There will be greater gains in repeating banner ads for consumers (1) who have low click-proneness coefficients (low click probability) in a given session and (2) who in general do not click much on average. Because additional banner ad exposures in prior sessions appear to increase the probability of a click in future sessions, there may be benefit to exposing consumers to banner ads even if they do not click on them during the current session. However, these gains will likely accrue only with repeat visits the site. Hence, while repeating banner ads in the short run (within a session) may lead to tedium effects that decrease beyond a threshold level of exposure, over the long run (across future sessions) repetition may increase click probability.

The level of banner ad exposure that leads to these gains varies across sessions for each consumer and across consumers. It appears that gains from repetition accrue earlier (and hence more) in sessions where consumers click more on banner ads in general. Click behavior in prior sessions predicts click behavior in the future with more clicks occurring when consumers are exposed to the banner ads again after a relatively long interval. Banner advertising exposures in prior sessions impacts banner exposure wearout in future sessions. Finally, we found that an increase in visit frequency appears to be associated with lower click probability in general. However, this implies that declining click-proneness across sessions at the site can be mitigated if consumers visit the site less frequently, or if the revisit is cycle is longer. Also note that our split-sample results indicated that an increase in visit frequency may not always be associated with lower click probability. Future modeling efforts will be necessary to examine and build on these results.

Our results also have relevance for two important issues in the online advertising industry: (1) declining click rates and (2) exposure-based versus performance-based pricing models.

In the long run, as the Web matures as a commercial medium, the results that more banner exposures in prior sessions led to higher click probabilities in the current session implies that the proportion of click-throughs generated through repetitions will increase. Because higher levels of banner exposures will be required to generate these gains, we believe click rates will continue to decline over time—a natural consequence as consumers gain experience using the medium and become more selective in how they allocate their cognitive resources and time online. Generating excessive number of exposures in each session may not be an option for some sites that attract goal-directed consumers and where consumers browse through few pages and then jump to other sites. In those cases, embedding ads within relevant content and using executions that arouse curiosity could be the key to inducing clicks.

Because clicks are most likely to occur during initial banner ad exposures, and consumers become less click-prone as they become more familiar with the site over time, our findings help support the current industry practice of higher prices for banners placed on entry and popular pages. Our research also suggests a theoretical and empirical basis to support impression-based pricing. Theoretically, our model suggests that banner exposures that do not immediately lead to a click may still lead to enduring communication outcomes. Empirically, we show that banner ad exposures in prior sessions have a significant positive effect on click probability in future sessions. Further, in situations where banner ads are placed across a network of sites, it is possible that banner ad exposures generated at one site may lead to clicks elsewhere. Under performance-based pricing models, Web sites that generate the cumulative effect—but not the click—are penalized. So it is not surprising that hybrid pricing models combining exposure and performance are becoming increasingly important in practice.

It is important to keep in mind that our model was built to predict click behavior over a relatively short period of time (i.e., less than a year), so this must be considered when evaluating its ability to explain the variance in future click-through rates over longer horizons.

The dynamic nature of the Internet market suggests that the model may have some difficulty predicting long-term trends in click behavior five or more years later. Modeling challenges in this arena include the continual innovation of new ad forms, a constantly changing landscape of Web sites that sell advertising

space, and changes in consumer response that are not based on consumer reaction to online advertising.

Thus, future modeling efforts may wish to take these effects into account as they attempt to capture long-term trends in click-through rates. Likely model variable candidates include terms that capture long-term changes in the structure of the online advertising market, such as the number and type of advertiser-supported Web sites, number and type of ad forms (including emerging alternatives to banner ads), and consumer behavior variables.

### 5.3. Concluding Remark

Our research marks one of the first attempts to model click outcomes from advertising exposure data, but click behavior is just one measure of advertising effectiveness. Increasingly, online advertisers are ultimately interested in responses subsequent to the click, such as purchase. *The first challenge for modeling such effects lies in more strategic data collection*. Data linking advertising exposure and subsequent market responses can only be collected through advertising networks and collaborative marketing alliances between sites and advertisers.

While privacy concerns and regulatory threats loom large (Tedeschi 2000), coming broadband access and improved measurement technology will soon make it possible to model the entire hierarchy of effects from advertising exposure and information search across several sites to product trial and purchase across several product categories. Availability of actual content consumption data will also make it possible to develop targeted ad delivery methods at the level of a paragraph on a Web page and enable the development of consumer profiles that tie content preferences directly to online purchase behavior. The availability of rich multisite consumer identified clickstream data will lead to more sophisticated modeling of consumer online behavior and more effective online marketing strategies.

## Acknowledgments

The authors acknowledge research support from the National Science Foundation from an NSF Economic, Decision and Management Science Grant (Number SBR 9422780) and thank an eLab (http://elab.vanderbilt.edu/) corporate sponsor, who wishes to remain anonymous for its generous support of this project. This research is based on the doctoral dissertation of the first author.

## Appendix A: Pooling Tests of Data

Because clicks on banner ads represent rare events, we tested for homogeneity to determine if pooling the data across the two sponsors would be appropriate. Pooling offers the major opportunity to gain degrees of freedom and improve reliability of estimates, but is appropriate only if the homogeneity hypothesis can be accepted.

We tested the equality of intercepts and slopes (overall homogeneity) for both sponsors on the basis of a comparison of the sum of the error sums of squares from the separate regressions with the error sum of squares of the pooled estimates using iterative generalized least squares procedure proposed by Gatignon and Reibstein (1986). The results appear in the following table. $RSS_2 + RSS_1 - R_w = 18.1$, whereas critical $\chi^2$ with 18 *df* at $\alpha = 0.05$ is 19.67. The test of equality of coefficients using the difference of residual sums of squares with 18 *df* indicates that the hypothesis of equal coefficients is not rejected. Therefore, pooling data for the two sponsors is appropriate.

**Iteratively Reweighted Least Square Logit Results (Homogeneity Tests)**

| Group | *RSS* | MSE | *df* | *N* |
|---|---|---|---|---|
| Unrestricted | | | | |
| Sponsor #15 | 66,855.8 | 2.789 | 23,956 | 23,974 |
| Sponsor #34 | 31,464.7 | 2.941 | 10,691 | 10,709 |
| | 98,320.5 | | 34,647 | |
| Completely restricted | 98,338.6 | 2.83 | 34,665 | 34,683 |

The homogeneity test statistic is $RSS_2 + RSS_1 - R_w \sim \chi^2_{df=K}$, where:

$RSS_1$ = Error sums of squares obtained from the regression for sponsor #15

$RSS_2$ = Error sums of squares obtained from regression for sponsor #34

$RSS_W$ = Error sums of squares obtained from the pooled regression

$K$ = number of parameters estimated ($K = 18$ in this research).

The $\chi^2$ test has $k$ degrees of freedom.

## Appendix B: Model Estimation

We specify a random effects estimator to model click outcomes at each banner exposure and as a function of covariates capturing browsing behavior at each exposure occasion, session, and individual's activity at the Web site. The outcome variable, whether a consumer $i$ clicked on a banner ad on exposure occasion $o_{is}$ in session $s_i$, is a Bernoulli event. Equations (4)–(7) specify a random-effects model to capture heterogeneity across consumers and change in response parameters across sessions while accounting for correlation in successive measures of click response variables. Because random effects are specified at two levels—systematic response heterogeneity (variation in response

coefficients across sessions) and random response heterogeneity (variation across consumers)—the random-parameters logit simulator by Train (1995) and Revelt and Train (1998) cannot be used.

Such models have been extensively studied in the multilevel modeling literature. Both likelihood approaches (MQL and PQL: marginal and penalized quasi-likelihood) and Bayesian methods (empirical or MCMC estimations using adaptive hybrid Metropolis-Gibbs sampling) have been suggested for estimating models with Bernoulli outcomes and nested or hierarchical random effects (see Rodriguez and Goldman 1995 for a detailed review). As Browne and Draper (2003) recommend in their comparison of Bayesian and likelihood methods for fitting random-effects logistic regression models, we use MLE for its computational speed during the model exploration phase and Bayesian estimation using MCMC to produce final publishable results with an appropriate diffuse prior.

For ease of exposition we respecify the random-effects logistic regression Equation (7) as

$$\begin{aligned}\pi_{iso} &= f\{\alpha^0 + \theta X_{iso} + \beta Y_{is} + \zeta_{is} + \eta_l\} \\ &= \{1 + \exp(-[\alpha^0 + \theta X_{iso} + \beta Y_{is} + \zeta_{is} + \eta_l])\}^{-1}, \end{aligned} \tag{A.1}$$

where $\zeta_{is} \sim N(0, \sigma_\zeta^2)$ and $\eta_l \sim N(0, \sigma_\eta^2)$ and the variables $X_{iso}$ and $Y_{is}$ are composite scales with the model containing many variables that vary across banner exposure occasions in a session and those that vary across sessions. Note that the random effects of within session variables and random effects of across session variables are correlated; however, we assume no correlations between random effects of within and across-session variables.

## Maximum Likelihood Estimation Approaches

The conditional likelihood function has the binomial form

$$L(\alpha, \theta, \beta \mid \zeta, \eta) = \prod_{i=1}^{I} \left( \frac{\pi_{iso}}{1 - \pi_{iso}} \right)^{Click_{iso}} (1 - \pi_{iso}). \tag{A.2}$$

To obtain the unconditional likelihood we need to multiply Equation (A.2) by the density of the random effects and integrate them out. However, this is intractable. Quasi-likelihood solutions address this by linearizing the exponential function in Equation (A.1) via a Taylor-series expansion so that it assumes the form of a standard normal model and then apply quasi-likelihood estimation using the binomial distribution assumption (details in Goldstein 1995).

We use a first-order Taylor expansion for the fixed part about the current estimates. For the second-order expansion for the random part we expand about zero, and we show below how this is modified to obtain improved estimates. We obtain at the $(t+1)$th iteration of the iterative generalized least squares (IGLS) algorithm (Goldstein 1995)

$$\begin{aligned} f(H_{t+1}) &= f(H_t) + X_{iso}(\hat{\theta}_{t+1} - \hat{\theta}_t) f'(H_t) + Y_{ij}(\hat{\beta}_{t+1} - \hat{\beta}_t) f'(H_t) \\ &\quad + \zeta_{is} f'(H_t) + \zeta_{is}^2 f''(H_t)/2, \end{aligned} \tag{A.3}$$

where $f'(H) = f(H)[1 + \exp(H)]^{-1}$, $f''(H) = f'(H)[1 - \exp(H)][1 + \exp(H)]^{-1}$. The term on the right-hand side of (A.3) updates the fixed part of the model and is equivalent to the standard iteratively reweighted least squares algorithm, which leads to MLE. The third and fifth term on the right-hand side of Equation (A.3) leads to first order adjustment (Goldstein 1995). Because (A.3) is essentially a linear model, procedures for linear hierarchical models can be used.

If $H_t = X_{iso}\hat{\theta}_t + Y_{is}\hat{\beta}_t$ it uses only the fixed part predictor for the Taylor expansion and is referred to as MQL (Breslow and Clayton 1993). However, $H_t = X_{iso}\hat{\theta}_t + Y_{is}\hat{\beta}_t + \hat{\zeta}_{is} + \hat{\eta}_s$ uses the Taylor expansion about the current estimated residuals, or posterior means of random effect and is referred to as penalized quasi-likelihood (PQL) (Breslow and Clayton 1993). While the PQL estimates are less biased compared to MQL due to their higher order of expansion, they lead to substantial downward bias when random effects are large. For this reason PQL estimates are used to generate starting values for the Bayesian estimation.

## Bayesian Approaches

To specify a Bayesian model, priors need to be placed on the macro parameters. Without strong intuition about the macro parameters, the priors are assumed diffuse. The particular structure is

$$u_i = \begin{pmatrix} \zeta_{is} \\ \eta_i \end{pmatrix} \overset{IID}{\sim} N_2(0, V_u), \quad V_u = \begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix}. \tag{A.4}$$

The particular structure for the prior is

$$\begin{aligned} \gamma &= \begin{pmatrix} \theta \\ \beta \end{pmatrix} \sim MVN(0, \Sigma_\gamma) \\ \Sigma_\gamma^{-1} &= \begin{matrix} 0.001 & 0 & 0 \\ 0 & 0.001 & 0 \\ 0 & 0 & 0.001. \end{matrix} \end{aligned} \tag{A.5}$$

We use $\chi^{-2}(\nu_\zeta, \sigma_\zeta^2)$ and $\chi^{-2}(\nu_\eta, \sigma_\eta^2)$ priors for random-effects variances $\sigma_\zeta^2$ and $\sigma_\eta^2$, respectively. The inclusion of these priors in Equation (7) leads to the full posteriors of $f(Click_{iso} \mid \theta, \beta, \zeta_{is}, \eta_l)$ as in Browne and Draper (2003) which we omit expressing in its entirety here because of its length.

This distribution does not lend itself easily to direct sampling. We use the Metropolis-Gibbs approach implemented in the software package MlwiN in which Gibbs sampling is used for variances and random-walk Metropolis sampling with Gaussian proposal distributions is employed for fixed effects and residuals. We use scaled versions of the estimated covariance matrices from PQL estimates to set the initial values of the proposal distribution variances.

We define $\varphi$ as the vector of all unknowns $(\alpha, \beta, \eta, \theta, \zeta, V_u)$ where each element of $\varphi$ is one of the elements of the unknown vectors. The joint posterior for $\varphi$ given the data and priors is the product of the conditional density of each of the elements of $\varphi$, given the true value of every other element of $\varphi$ (which are unknown). We construct a Markov chain whose stationary distribution is equivalent to the posterior distribution of $\varphi$. Let $\varphi_n^i$ represent the $i$th

iteration of unknown $n$, where the vector $\theta^0$ is the starting value of the chain. The chain then iterates in the following manner by drawing successively from the distributions.

In step 1, fixed effect coefficients $\theta$ is updated using the random-walk Metropolis as follows: For $l = 0, \ldots, m$ and with $\theta_{(-m)}$ signifying the $\theta$ vector without component $m$

$$\theta_m^{(t)} = \theta_m^* \quad \text{with probability } \min\left[1, \frac{p(\theta_m^* \mid c, \zeta, \eta, \theta_{(-m)})}{p(\theta_m^{(t-1)} \mid c, \zeta, \eta, \theta_{(-m)})}\right]$$
$$= \theta_{\mathrm{m}}^{9(t-1)} \quad \text{otherwise,}$$

where

$$\theta_m^* \sim N(\theta_{\mathrm{m}}^{9(t-1)}, \sigma_{1m}^2),$$

and

$$p(\theta_m \mid c, \zeta, \eta, \theta_{(-m)}) \propto \prod_{iso}\left[1 + e^{-(X\theta)_{iso}}\right]^{-Click_{iso}}\left[1 + e^{(X\theta)_{iso}}\right]^{Click_{iso}-1}. \quad \text{(A.6)}$$

In step 2, Metropolis-updates the across-session residuals $\zeta_{is}$: for $s = 1, \ldots, s_i$ and $i = 1, \ldots, I$

$$\zeta_{is}^{(t)} = \zeta_{is}^* \quad \text{with probability } \min\left[1, \frac{p(\zeta_{is}^* \mid C, \eta, \theta, \sigma_\zeta^2)}{p(\zeta_{is}^{(t-1)} \mid C, \theta, \sigma_\zeta^2)}\right]$$
$$= \zeta_{is}^{(t-1)} \quad \text{otherwise,}$$

where

$$\zeta_{is}^* \sim N(\zeta_{is}^{(t-1)}, \sigma_{2is}^2),$$

and

$$p(\zeta_{is} \mid c, \eta, \theta, \sigma_\zeta^2) \propto \sigma_\zeta^{-1}\exp\left(-\frac{\zeta_{is}}{2\sigma_\zeta^2}\right) \cdot \prod_{iso}\left[1 + e^{-(X\theta)_{iso}}\right]^{-Click_{iso}}\left[1 + e^{(X\theta)_{iso}}\right]^{Click_{iso}-1}. \quad \text{(A.7)}$$

In step 3, a Metropolis update of across-consumer residuals $\eta_I$

$$\eta_i^{(t)} = \eta_i^* \quad \text{with probability } \min\left[1, \frac{p(\eta_i^* \mid C, \zeta, \theta, \sigma_\eta^2)}{p(\eta_i^{(t-1)} \mid C, \zeta, \theta, \sigma_\eta^2)}\right]$$
$$= \eta_i^{(t-1)} \quad \text{otherwise,}$$

where

$$\eta_i^* \sim N(\eta_i^{(t-1)}, \sigma_{3i}^2),$$

and

$$p(\eta_i \mid c, \zeta, \theta, \sigma_\eta^2) \propto \sigma_\eta^{-1}\exp\left(-\frac{\eta_i}{2\sigma_\eta^2}\right) \cdot \prod_{iso}\left[1 + e^{-(X\theta)_{iso}}\right]^{-Click_{iso}}\left[1 + e^{(X\theta)_{iso}}\right]^{Click_{iso}-1}. \quad \text{(A.8)}$$

In step 4, Gibbs update of across-session variance $\sigma_\zeta^2$

$$(\sigma_\zeta^2 \mid \zeta) \sim V^{-1}\left[\frac{N_J + \nu_\zeta}{2}, \frac{1}{2}\left(\nu_\zeta S_\zeta^2 + \sum_{is}\zeta_{jk}^2\right)\right], \quad \text{(A.9)}$$

where $N_J = \sum_{k=1}^{K} J_k$ is the total number of level 2 units.

Finally, step 5 is a Gibbs update of level 3 variance $\sigma_\eta^2$

$$(\sigma_\eta^2 \mid \eta) \sim V^{-1}\left[\frac{I + \nu_\eta}{2}, \frac{1}{2}\left(\nu_\eta s_\eta^2 + \sum_i \eta^2\right)\right],$$

where $I$ is the number of consumers in the sample. We derive the specification of the variances $\sigma_{1m}$, $\sigma_{2is}$, and $\sigma_{13i}$, of the proposal distributions from the PQL estimates.

## References

Allenby, Greg M., Peter E. Rossi. 1999. Marketing models of consumer heterogeneity. *J. Econom.* **89** 57–78.

Ariely, Dan. 2000. Controlling the information flow: Effects on consumers' decision making and preferences. *J. Consumer Res.* **27**(2) 233–248.

Batra, Ravi, Michael L. Ray. 1986. Situational effects of advertising repetition: The moderating influence of motivation, ability and opportunity to respond. *J. Consumer Res.* **12** 432–445.

Ben-Akiva, Moshe, Steven R. Lerman. 1985. *Discrete Choice Analysis*. MIT Press, Cambridge, MA.

Benway, J. P. 1998. Banner blindness: The irony of attention grabbing on the World Wide Web. *Proc. Human Factors and Ergonomics Soc. 42nd Annual Meeting* **1** 463–467.

Berlyne. 1970. Novelty, complexity, and hedonic value. *Perception Psychophysics* **8** 279–286.

Bettman, James. 1979. *An Information Processing Theory of Consumer Choice*. Addison Wesley, Reading, MA.

Bicknell, Craig. 1999. Net ad rates continue to fall. *Wired News* (January 25). http://www.wired.com/news/business/story/17520.html.

Breslow, N. E., D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. *J. Amer. Statist. Association* **88** 9–25.

Briggs, Rex, Nigel Hollis. 1997. Advertising on the Web: Is there response before click-through? *J. Advertising Res.* **37**(2) 33–45.

Browne, W. J., D. Draper. 2003. A comparison of Bayesian and likelihood methods for fitting multilevel models. *Bayesian and Likelihood Methods in Multilevel Modeling*. Forthcoming.

Bryk, A. S., S. W. Raudenbush, R. T. Congdon. 1996. *Hierarchical Linear and Nonlinear Modelling with the HLM12L and HLM13L Programs*. Scientific Software International, Inc., Chicago, IL.

Buchanan, Bruce, Donald G. Morrison. 1985. A stochastic model of list falloff with implications for repeat mailings. *J. Direct Marketing* **2**(3) 7–14.

Calder, Bobby J., Brian Sternthal. 1980. Television commercial wearout: An information processing view. *J. Marketing Res.* **17**(May) 173–186.

Chatterjee, Patrali. 1998. Modeling consumer network navigation at World Wide Web sites—Implications for advertising. Doctoral dissertation, Vanderbilt University, Nashville, TN.

Chen, Yuxin, Ganesh Iyer. 2002. Consumer addressability and customized pricing. *Marketing Sci.* **21**(2) 197–208.

Chintagunta, Pradeep K., Dipak C. Jain, Naufel J. Vilcassim. 1991. Investigating heterogeneity in brand preferences in logit models for panel data. *J. Marketing Res.* **28**(November) 417–428.

*DoubleClick*. 1996. Frequency and Banner Burnout. http://www.doubleclick.net/nf/general/frequset.htm.

Dreze, Xavier, Francois Hussherr. 2003. Internet advertising: is anybody watching. *J. Interactive Marketing* **17**(4). Forthcoming.

*eMarketer*. 2002. *Essential e-Business Numbers for Marketers, Q3*. October, New York.

——. 2001. *The eAdvertising Report*. April, New York.

Gatignon, H., D. Reibstein. 1986. Pooling logit models. *J. Marketing Res.* **23**(August) 281–285.

Goldstein, Harvey. 1995. *Multilevel Statistical Models*, 2nd ed. Edward Arnold, London.

Greenwald, Anthony G., Clark Leavitt. 1984. Audience involvement in advertising: Four levels. *J. Consumer Res.* **11**(June) 581–592.

Hanson, Ward. 2000. *Internet Marketing*. South-Western Publishing, Cincinnati, OH.

Hoffman, Donna L., Thomas P. Novak, Marcos A. Peralta. 1999. Building consumer trust in online environments: The case for information privacy. *Comm. ACM* **42**(4) 80–85.

——, ——. 2000. Acquiring customers on the Web. *Harvard Bus. Rev.* (May–June) 179–188.

——, ——. 1996. Marketing in hypermedia computer-mediated environments: Conceptual foundations. *J. Marketing* **60**(July) 50–68.

Lynch, John G., Dan Ariely. 2000. Wine online: Search costs affect competition on price, quality and distribution. *Marketing Sci.* **19**(1) 83–103.

Huberman, Bernado A., Peter L. T. Pirolli, James E. Pitkow, Rajan M. Lukose. 1998. Strong regularities in World Wide Web surfing. *Science* **280**(5360) 95–97.

Janiszewski, Chris. 1993. Preattentive mere exposure effects. *J. Consumer Res.* **20**(December) 376–392.

Jones, J. Morgan, Jane T. Landwehr. 1988. Removing heterogeneity bias from logit model estimation. *Marketing Sci.* **7**(1) 41–59.

Jupiter Research. 2003. Jupiter Research Internet Advertising Model, 7/03 (US only).

King, Gary, Langche Zeng. 2000. Logistic regression in rare events data. *Political Anal.* **92** 1–27.

MacInnis, Deborah J., Christine Moorman, Bernard J. Jaworski. 1991. Enhancing and measuring consumers' motivation, opportunity, and ability to process brand information from ads. *J. Marketing* **55**(October) 32–53.

Mitchell, Andrew A. 1983. Cognitive processes initiated by exposure to advertising. R. Harris, ed. *Information Processing Research in Advertising*. Lawrence Erlbaum Associates, Hillsdale, NJ, 13–42.

Morrison, Donald G. 1969. On the interpretation of discriminant analysis. *J. Marketing Res.* **6**(May) 156–163.

Novak, T. P., D. L. Hoffman, Y. F. Yung. 2000. Measuring the customer experience in online environments: A structural modeling approach. *Marketing Sci.* **19**(1) 22–44.

Pechmann, Cornelia, David W. Stewart. 1989. Advertising repetition: A critical review of wearin and wearout. James H. Leigh, R. Martin, Jr., eds. *Current Issues and Research in Advertising 1988*. Claude University of Michigan, Ann Arbor, MI, 285–329.

Peterman, Michelle L., Harper A. Roehm, Jr., Curtis P. Haugtvedt. 1999. An exploratory attribution analysis of attitudes toward the World Wide Web as a product information source. *Adv. Consumer Res.* **26** 75–79.

Pieters, Rik, Edward Rosbergen, Michel Wedel. 1999. Visual attention to repeated print advertising: A test of scanpath theory. *J. Marketing Res.* **36**(December) 305–314.

Revelt, D., K. Train. 1998. Mixed logit with repeated choices: Households choices of appliance efficiency level. *Rev. Econom. Statist.* **80**(4) 647–657.

Rodriguez, G., N. Goldman. 1995. An assessment of estimation procedures for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser.* **A**(158) 73–89.

Rosbergen, Edward, Rik Pieters, Michel Wedel. 1997. Visual attention to advertising: A segment-level analysis. *J. Consumer Res.* **24**(December) 305–314.

Schroeder, Will. 1998. Testing Web sites with eye-tracking. *User Interface Engineering—Eye for Design* **5**(5) (Sept.–Oct.) 3–4.

Sen, Shahana, Balaji Padmanabhan, Alexander Tuzhilin, Norman H. White, Roger Stein. 1998. The identification and satisfaction of consumer analysis-driven information needs of marketers on the WWW. *Eur. J. Marketing* **32**(7–8) 688–702.

Sohn, Dongyoung, John D. Leckenby. 2001. Locus of control and interactive advertising. Paper presented at the 2001 Annual Conference of the American Academy of Advertising, Salt Lake City, Utah, March. http://www.ciadvertising.org/studies/reports/info_process/locus.htm.

Tedeschi, Bob. 2000. DoubleClick puts off plan for wider use of personal data. *New York Times on the Web*, March 2, Business Desk. www.nytimes.com/.

Train, K. E. 1998. Recreation demand models with taste variation over people. *Land Econom.* **74**(2) 230–239.

*This paper was received May 26, 1998, and was with the authors 41 months for 5 revisions; processed by Scott Neslin.*