ANOCHA ARIBARG, RIK PIETERS, and MICHEL WEDEL*

Advertising recognition tests use advertisements as visual retrieval cues; they require consumers to report which advertisements they remember having seen earlier and whether they noticed the advertised brand and read most of the text at the time. Using a heterogeneous randomly stopped sum model, the authors establish the relationship between consumers' actual attention to print advertisements, as measured through eye tracking, and subsequent ad recognition measures. They find that ad recognition measures are systematically biased because consumers infer prior attention from the ad layout and their familiarity with the brands in the advertisements. Such biases undermine the validity of recognition tests for advertising practice and theory development. The authors quantify the positive and negative diagnostic value of ad recognition for prior attention and demonstrate how these diagnostic values can be used to develop bias-adjusted recognition (BAR) scores that more accurately reflect prior attention. Finally, the authors show that differences in the scores from ad recognition tests based on in-home versus lab exposure attenuate when the bias-adjustment procedure is applied.

Keywords: recognition test, advertising, bias correction, stopped sum model, Bayesian

Raising the BAR: Bias Adjustment of Recognition Tests in Advertising

In 1923, Daniel Starch (see also Shepard 1942) pioneered ad recognition tests, and they have been used ever since in marketing. In these tests, consumers report which advertisements they remember having seen at an earlier time when they were exposed to a specific magazine (the "ad-noted" measure), whether they identified the advertised brand (the "brand-associated" measure), and whether they read most of the copy in the advertisement (the "read-most" measure). Ad recognition tests provide measures of consumers' direct memory for prior exposure to advertising. The more attention consumers pay to the advertisement, brand, and text, the higher the recognition scores in question are assumed to be. Though originally developed for print advertisements, recognition tests are also used to assess prior exposure to outdoor (Bhargava, Dontu, and Caron 1994), Web (Havlena and Graham 2004), and television (Heath and Nairn 2005; Mehta and Purvis 2006; Singh, Rothschild, and Churchill 1988) advertising, among others. Recognition scores have been popular metrics of ad effectiveness in advertising practice, in which recognition is assessed after participants have been exposed to advertisements in their homes (Baldinger and Cook 2006; Belch and Belch 2001; Hanssens and Weitz 1980). Such metrics are also frequently used for testing ad processing in academic advertising research, either from secondary data (Finn 1988) or under more controlled laboratory conditions (Mothersbaugh, Huhmann, and Franke 2002; Puntoni and Tavassoli 2007). Ad recognition tests are easy to administer, with advertisements used as retrieval cues, and the resultant scores are readily comparable to benchmarks based on a long history of applications. These strengths contribute to their popularity.

Despite the extensive application of recognition tests, little is known about their accuracy as measures of attention to advertisements during prior exposure. This is surprising because memory research indicates that ad recognition may

*Anocha Aribarg is Assistant Professor of Marketing, Stephen M. Ross School of Business, University of Michigan (e-mail: anocha@umich.edu). Rik Pieters is Professor of Marketing, Department of Marketing, Faculty of Economics and Business Administration, University of Tilburg (e-mail: pieters@uvt.nl). Michel Wedel is Pepsico Professor of Consumer Research, Robert H. Smith School of Business, University of Maryland (e-mail: mwedel@rhsmith.umd.edu). Christophe Van den Bulte served as associate editor for this article.

ISSN: 0022-2437 (print), 1547-7193 (electronic)

Journal of Marketing Research
Vol. XLVII (June 2010), 387–400

be systematically biased because of memory reconstruction processes during retrieval (Mitchell and Johnson 2000; Roediger and McDermott 2000; Yonelinas 2002). This argument casts doubts on the diagnostic value of recognition tests as a measure of consumers' prior attention to advertising and, thus, on their validity in gauging ad effectiveness and in developing advertising theory. Although prior research has established links between indirect memory measures and visual attention to advertisements (Wedel and Pieters 2000), often measured as gaze duration (Pieters and Wedel 2004), tests of the diagnostic value of Starch-type recognition tests for prior attention to advertising are unavailable. In addition, little is known about the stability of these recognition measures across the different exposure conditions used in academic research and in practice, which makes it challenging to generalize the findings obtained under lab conditions to in-home situations. Given the prevalent use of recognition testing by marketing academics to test advertising processing models and in marketing practice to assess which advertisements attract the most attention and to guide advertising message and media decisions (Hermie et al. 2005), we believe that it is imperative to address these research questions.[1]

This research makes the following three contributions: First, we propose a new statistical model to examine the relationship between attention to print advertisements, as measured through eye-tracking methodology, and Starch ad recognition measures. The model accommodates the potential influence of ad layout familiarity with the advertised brands on attention and recognition memory. As hypothesized, we observe that ad layout and brand familiarity indeed systematically bias ad recognition measures, independent of their effects on attention during the earlier ad exposure.

Second, informed by the literature on diagnostic testing in medical decision making, and based on the model, we quantify the diagnostic value of ad recognition measures for prior attention to advertising. We use Bayes' theorem to establish positive diagnostic values (PDVs) as the probability that consumers have actually seen a specific advertisement and its elements, given that they claim recognition, and we establish negative diagnostic values (NDVs) as the probability that consumers have actually not seen a specific advertisement and its elements, given that they do not claim recognition. Significant differences in PDV and NDV across different recognition measures are revealed. In particular, the ad-noted measure has a high PDV, and the brand-associated measure has a high NDV.

Third, we demonstrate how the PDV and NDV of ad recognition measures can be used to develop bias-adjusted recognition (BAR) scores. Bias adjustment may be particularly useful if eye-tracking measures of attention to the advertisements are not available. Holdout validation tests show that bias adjustment substantially improves the diagnostic value of ad recognition measures. We assess the stability of recognition measures across in-home and laboratory conditions and apply the bias-adjustment procedure to recognition scores in both conditions. The results reveal that our procedure helps mitigate the differences in the measures obtained from these two conditions. In the next section, we describe the data on which the analyses are based.

[1]See http://www.time.com/time/mediakit/audience/research/proprietary/starch.html (accessed March 2009).

DATA

Data were collected in cooperation with the market research agency Verify International (the Netherlands). Four hundred twenty-eight consumers (50% women, ages 18–60) participated in the study for monetary compensation. Two hundred forty-three randomly selected consumers from the participant pool of the market research agency received, at home, a copy of the latest issue of *Cosmopolitan* magazine, which contained all 48 full-page advertisements; they were asked to use the magazine as they normally would and to come to the market research agency's lab one week later for the ad recognition test. This situation mimics ad recognition testing in practice, though in practice, the time delay between exposure and testing varies (the Printed Advertising Rating Methods study investigates the effect of the time delay and finds modest effects of the delay for recognition; Bagozzi and Silk 1983).

The remaining 185 consumers were directly invited to the lab, and data were collected from them in three phases (which we describe subsequently). These participants were exposed to the same issue of *Cosmopolitan*, and their eye movements were recorded to obtain measures of attention to advertising. The ad recognition test was also subsequently administered to them. Participants in the in-home condition engaged in the same ad recognition test as those in the lab condition (Phase 3). All participants were not a priori made aware of the ad recognition test in Phase 3. Participants were not screened, except to ensure that they did not have abnormal vision.

Brand Familiarity

In Phase 1, participants provided general information about their sociodemographics and their familiarity with a large set of products and brands (total n = 91), as well as several other unrelated issues (e.g., media consumption). Participants were seated behind a touch-sensitive computer screen and were asked about brand familiarity: "You will see a number of brand names, please indicate how well-known each brand is to you." Participants responded to each brand name as follows: "completely unknown" (score = 0), "unknown" (1), "known" (2), and "known very well" (3).

Eye Tracking

In Phase 2, we assessed attention to advertising with eye tracking (Wedel and Pieters 2007). After a brief warm-up task, participants paged through a digital copy of the most recent issue of *Cosmopolitan* (containing the 48 full-page advertisements) in fixed front-to-back order while their eye-movements were recorded. They could inspect pages more closely if wanted, as when exploring a magazine at home (Janiszewski 1998), and pages could even be skipped entirely. All participants had normal or corrected-to-normal vision and had not participated in eye-tracking research before. None had seen the issue before. Instructions and stimuli were presented on NEC 21-inch LCD monitors in full-color bitmaps with a 1280 × 1024-pixel resolution. Participants touched the lower-right corner of the (touch-sensitive) screen to proceed, as when leafing through print material.

We used infrared corneal reflection methodology for eye tracking (Duchowski 2003). During data collection, participants could freely move their heads in a virtual box of approximately 30 centimeters while cameras tracked the position of the eye and head, allowing continuous correction of position shifts. Eye movements consist of sequences of saccades and fixations, periods during which the eye is relatively still and information uptake occurs. The duration of an individual fixation is approximately 200–400 milliseconds (Rayner 1998). Gaze duration is the sum of individual fixation durations on an advertisement or its elements; both fixation frequencies and gaze durations on the advertisement and its elements are common metrics of visual attention (Wedel and Pieters 2007). Fixation frequencies and gaze durations on the text, pictorial, and brand (logo; brand name in headline, slogan, or body text) as the main ad design elements were retained for each of the 185 participants and the 48 advertisements studied.

Ad Recognition

In Phase 3, participants were exposed to each of the target advertisements from *Cosmopolitan* on a computer screen (after verifying that they remembered having seen this issue of the magazine; all had), and they were asked to indicate for each advertisement, "When you went through this issue of *Cosmopolitan* …" (1) "Have you read or seen something of this specific advertisement?" (ad-noted: yes = 1, no = 0); if yes, (2) "Have you seen or read which brand was advertised?" (brand-associated: yes = 1, no = 0), and (3) "Have you read half (50%) or more of the text in the advertisement?" (read-most: yes = 1, no = 0). These are the three standard questions in Starch ad recognition tests (Finn 1988, 1992) and are similar to other ad recognition measures in ad theory and practice (Heath and Nairn 2005; Krishnan and Chakravarti 1999; Mehta and Purvis 2006). All advertisements were shown with their editorial facing page and in the order in which they appeared in the magazine. The test procedure was as similar as possible to a standard "through-the-book" procedure, in which the entire magazine, with editorial content and advertisements, is shown during the test. On completion, participants were debriefed (none indicated having expected the memory task when participating in the earlier phases of the study), thanked, and paid. Table 1 provides summary statistics.

As we expected, ad recognition scores, as a percentage of participants answering "yes" to each of the measures, differed between the lab and the in-home conditions. On average, 39.2% of participants in the in-home condition indicated recognition of the advertisements, compared with 54.3% in the lab condition ($p < .05$). In addition, the brand-associated score was 29.5% for participants in the in-home condition, compared with 40.5% in the lab ($p < .10$). Unlike the ad-noted and brand-associated scores, scores for the read-most measure were close for participants in the in-home (16.9%) and the lab (16.3%) conditions ($p > .10$).

Ad Content Analysis

Additional information about the advertisements, products, and brands was collected through content analysis. A panel of 20 trained coders (10 male and 10 female graduate students) judged the advertisements and brands in individualized random order on eight seven-point rating scales. Scores were averaged per advertisement across judges (average alpha for the 20 coders was .892 across the eight items). A principal components analysis on the eight ratings across the 48 target advertisements produced three clean

Table 1
DESCRIPTIVE STATISTICS

Variable	*N*	*M*	*SD*	*Mdn*	*Minimum*	*Maximum*
Advertisements: Surface Sizes						
Brand (inch²)	48	11.134	8.442	8.168	1.825	45.752
Pictorial (inch²)	48	64.947	16.229	68.536	8.791	81.632
Text (inch²)	48	15.499	11.575	16.792	.000	46.938
Laboratory Group (n = 185)						
Brand familiarity (0, ..., 4)	8880	1.873	.983	2	0	3
Fixation Frequency						
Brand (0, ..., n)	8880	2.824	3.296	2	0	38
Pictorial (0, ..., n)	8880	5.876	4.712	5	0	49
Text (0, ..., n)	8880	3.872	5.782	2	0	87
Total (0, ..., n)	8880	12.572	10.551	10	0	124
Gaze Duration						
Brand (seconds)	8880	.605	.765	.38	0	9.12
Pictorial (seconds)	8880	1.173	1.117	.86	0	14.48
Text (seconds)	8880	.811	1.303	.34	0	17.02
Total (seconds)	8880	2.589	2.469	1.92	0	26.22
Recognition Memory						
Ad noted (0, ..., 1)	8880	.543	.498			
Brand associated (0, ..., 1)	8880	.405	.491			
Read most (0, ..., 1)	8880	.163	.369			
In-Home Group (n = 243)						
Recognition Memory						
Ad noted (0, ..., 1)	11,664	.392	.488			
Brand associated (0, ..., 1)	11,664	.295	.456			
Read most (0, ..., 1)	11,664	.169	.375			

Notes: Mean values of recognition memory measures are proportions.

components (with eigenvalues >1): brand popularity, ad uniqueness, and ad attractiveness. We use mean orthogonal component scores across items in the three scales in the post hoc analyses. Brand popularity comprised three items: "I know this brand …" (1 = "not at all," and 7 = "very well"), "I have seen this specific advertisement for this brand …" (1 = "never before," and 7 = "very often"), and "I am … with this brand" (1 = "not at all familiar," and 7 = "very familiar"). Ad uniqueness comprised three items: "To me, this specific advertisement for this brand is …" (1 = "not at all unique," and 7 = "very unique"; 1 = "not at all original," and 7 = "very original"; 1 = "not at all unexpected," and 7 "very unexpected"). Finally, ad attractiveness comprised two items (1 = "not at all attractive," and 7 = "very attractive"; 1 = "not at all exciting," and 7 = "very exciting"). In addition, the number of words in the headline was counted (M = 4.67, SD = 2.71) because of its potential influence on attention to advertising (Rayner et al. 2001).

A MODEL OF ATTENTION AND AD RECOGNITION

We propose a model that specifies the relationship between attention to advertisements and subsequent ad recognition measures, and we use this to derive the diagnostic value of ad recognition tests for prior attention to advertisements. We calibrate the model on attention and ad recognition measures obtained from the 185 participants in the lab condition.

Attention Model

We have $l = 1, \ldots, L$ advertisements, each consisting of $j = 1, \ldots, J$ ad design elements; a sample of $i = 1, \ldots, I$ consumers; and $m = 1, \ldots, M$ recognition measures. There are J = 3 ad design elements—pictorial, text, and brand—and M = 3 recognition measures—ad-noted, brand-associated, and read-most. The data available for calibrating the model consist of the gaze duration of consumer i on element j of ad l ($t_{i,j,l}$) and the fixation frequency of consumer i on element j of ad l ($n_{i,j,l}$). The proposed attention component describes gaze duration as the sum of individual fixation durations through a hierarchical randomly stopped sum Poisson model. This model captures the mechanism through which gaze duration arises more accurately than previous research (Janiszewski 1998; Pieters and Wedel 2004; Wedel and Pieters 2000). A stopped sum distribution is defined as the distribution of the sum of $i = 1, \ldots, n$ independent and identically distributed random variables X_i, where n is the realization of a random variable N. The distribution of N is referred to as the sum distribution (in our case, a Poisson distribution), and the distribution of the X_i is referred to as the elementary distribution (in our case, an exponential distribution) (Johnson, Kotz, and Balakrishnan 1994; Stuart and Ord 1994). Thus, we model gaze duration on a specific element as the sum of the durations of the individual fixations on that element: $t_{i,j,l} = \sum_{k_i=1}^{n_{i,j,l}} d_{k_i,j,l}$, where $d_{k_i,j,l}$ is duration of the kth fixation of the ith individual. This defines the distribution of $t_{i,j,l}$ as a randomly stopped sum. The specification of the model is facilitated by writing the joint density of fixation frequency and gaze duration as the product of the marginal distribution of fixation frequency and the conditional distribution of gaze duration given fixation frequency (Heller et al. 2007). We assume that the marginal distribution of fixation frequency ($n_{i,j,l}$) is Poisson (Wedel and Pieters 2000) with parameter $\mu_{i,j,l}$ and the distribution of fixation durations ($d_{k_i,j,l}$) is exponential with parameter $\lambda_{i,j,l}$ (Harris et al. 1988). Conditional on $n_{i,j,l}$, $t_{i,j,l}$ is a sum of identically distributed exponential random variables with parameters $\lambda_{i,j,l}$ and thus follows a gamma distribution (Johnson, Kotz, and Balakrishnan 1994), with parameters $n_{i,j,l}$ and $\lambda_{i,j,l}$ and expectation $n_{i,j,l}\lambda_{i,j,l}$. Thus, we have the following:

$$
\begin{aligned}
(1)\qquad \text{Fixations: } f_N\left(n_{i,j,l}\right) &= \text{Poisson}\left(n_{i,j,l}\middle|\mu_{i,j,l}\right)\\
&= \frac{\mu_{i,j,l}^{n_{i,j,l}}\exp\left[-\mu_{i,j,l}\right]}{n_{i,j,l}!}, \text{ and}\\
\text{Gaze: } f_{T|N}\left(t_{i,j,l}\middle|n_{i,j,l}\right) &= \text{gamma}\left(t_{i,j,l}\middle|n_{i,j,l},\lambda_{i,j,l}\right)\\
&= t_{i,j,l}^{n_{i,j,l}-1}\frac{\left[\exp -t_{i,j,l}/\lambda_{i,j,l}\right]}{\lambda_{i,j,l}^{n_{i,j,l}}\Gamma\left(n_{i,j,l}\right)}.
\end{aligned}
$$

Expected fixation frequency, $\mu_{i,j,l}$, is parameterized as a function of explanatory variables (but not the expected fixation duration $\lambda_{i,j,l}$ because it is largely beyond cognitive control and essentially random; Harris et al. 1988):

$$
\begin{aligned}
(2)\qquad \mu_{i,j,l} &= \exp\left(x_{i,j,l}^{A\prime}\alpha_{i,j}\right), \text{ and } \alpha_{i,1:J} \sim \text{MVN}\left(\bar{\alpha}, D_\alpha\right),\\
\lambda_{i,j,l} &= \exp\left(\lambda_{i,j,l}^{*}\right), \text{ and } \lambda_{i,1:J}^{*} \sim \text{MVN}\left(\bar{\lambda}, D_\lambda\right),
\end{aligned}
$$

where $\alpha_{i,1:J} \equiv \text{vec}(\alpha_i)$, with α_i being a $(J \times P^A)$ matrix, where P^A is the dimension of explanatory variables (including the intercept) $x_{i,j,l}^{A}$, and $\lambda_{i,1:J}^{*} = (\lambda_{i,1}^{*}, \lambda_{i,2}^{*}, \lambda_{i,3}^{*})'$. These parameters follow multivariate normal distributions, as we show in Equation 2, to account for heterogeneity among consumers and overdispersion of the fixation counts. Thus, this model component extends the multivariate Poisson lognormal distribution (Chib and Winkelmann 2001). In Equation 2, we account for the influence of the ad layout (as a stimulus-related factor) on fixation frequency in terms of the sizes of the brand, pictorial, and text elements. That is, larger surface sizes enhance figure–ground segmentation and increase the salience of ad elements (Itti 2005), which should increase attention to them (Pieters and Wedel 2004; Wedel and Pieters 2000). Brand familiarity (as a person-related factor) is also predicted to influence attention to the advertisement and its elements (Reichle, Rayner, and Pollatsek 2003). Therefore, we include all these variables in $x_{i,j,l}^{A}$ in Equation 2. The parameters $\alpha_{i,1:J}$ reflect the direct effects of these variables on attention.

Recognition Memory Model

The binary variables indicate a "yes" or "no" response for recognition measure m for consumer i for ad l, $y_{i,m,l}$. The recognition memory component is a two-stage multivariate probit model (Edward and Allenby 2003; Manchanda, Ansari, and Gupta 1999), in which attention is specified to affect multiple correlated memory measures. The two-stage model reflects the structure of the recognition questions: First, a person indicates whether he or she remembers seeing the advertisement and, if so, whether he or she remembers having identified the brand and having read most of the text. Attention is assumed to be unobserved but reflected in the total gaze duration (Rayner 1998). Recognition occurs when the strength of the memory signal, which is a function of prior attention, exceeds a threshold (Hintzman 2000). The attention and memory components of the model both allow for unobserved heterogeneity among individuals and are estimated simultaneously.

Recognition memory for consumer i, ad l, and recognition measure m are as follows:

$$\text{(3)}\quad \text{Ad-noted: } f_Y\left(y_{i,1,l}\middle|\omega_{i,1,l}\right) = P\left(y_{i,1,l}=1\middle|\omega_{i,1,l}\right)^{y_{i,1,l}} P\left(y_{i,1,l}=0\middle|\omega_{i,1,l}\right)^{1-y_{i,1,l}},$$

$$\text{Brand-associated and read-most: } f_Y\left(y_{i,m,l}\middle|y_{i,1,l}=1,\omega_{i,\mathbf{m},l}\right) = P\left(y_{i,1,l}=1\middle|\omega_{i,1,l}\right)P\left(y_{i,m,l}=1\middle|y_{i,1,l}=1,\omega_{i,1,l}\right)^{y_{i,m,l}} P\left(y_{i,m,l}=0\middle|y_{i,m,l}=1,\omega_{i,m,l}\right)^{1-y_{i,m,l}},\ m=2,3.$$

Expected memory $\omega_{i,m,l}$ is parameterized as a function of explanatory variables:

$$\text{(4)}\quad \omega_{i,1,l} = \beta_{i,1,0} + \phi_{i,l}\beta_{i,1,\phi} + x^{M\prime}_{i,1,l}\beta_{i,1}, \text{ and}$$

$$\beta_i \equiv \left(\beta_{i,1,0}, \beta'_{i,1,\phi}, \beta'_{i,1}\right)' \sim \text{MVN}\left(\bar{\beta}, D_\beta\right)$$

$$\omega_{i,m,l} = \gamma_{i,m,0} + \phi_{i,\ell}\gamma_{i,m,\phi} + x^{M\prime}_{i,m,l}\gamma_{i,m}, \text{ and}$$

$$\gamma_i \equiv \left(\gamma_{i,2,0}, \gamma'_{i,2,\phi}, \gamma'_{i,2}, \gamma_{i,3,0}, \gamma'_{i,3,\phi}, \gamma'_{i,3}\right)' \sim \text{MVN}\left(\bar{\gamma}, D_\gamma\right),\ m=2,3,$$

where the explanatory variables are $\phi_{i,l}$, attention to each of the three ad elements (as we explain in detail subsequently), and $x^M_{i,m,l}$, the size of the ad elements and brand familiarity. The parameters β_i and γ_i follow multivariate normal distributions, as we show in Equation 4, to account for heterogeneity among consumers. Note that we assume that the individual-level parameters are uncorrelated across Equations 2 and 4.[2]

We model the probability that a consumer claims to have noted the advertisement (m = 1), identified the brand (m = 2), and read most of the text (m = 3) as a function of attention to the advertisement and the ad elements in question. Attention is reflected in fixation frequency and fixation duration (Rayner 1998) and, therefore, in the total gaze duration. Yet gaze duration is not a perfect indicator of unobserved attention (Pieters and Wedel 2007). For example, Henderson (1992) describes the relationship through a "rubber-band" metaphor, with the eyes and attention closely but imperfectly coupled. Therefore, we assume that gaze duration on an ad element for a specific advertisement is an unbiased but imprecise indicator of attention to that ad element. Attention to an element is operationalized as the expected gaze duration: $E[n_{i,j,l}]E[t_{i,j,l}|n_{i,j,l}] = \mu_{i,j,l}\lambda_{i,j,l}$. We assume that each of the three memory measures can be affected by attention to each of the three ad elements, so that $\phi_{i,l} = (\mu_{i,1,l}\lambda_{i,1,l}, \mu_{i,2,l}\lambda_{i,2,l}, \mu_{i,3,l}\lambda_{i,3,l})$ in Equation 4, the (3 × 1) parameter vector $\beta'_{i,1,\phi}$, contains the individual-specific attention weights, capturing the effects of attention on ad-noted. Similarly, $\gamma'_{i,j,\phi}$ captures the effects of attention on brand-associated (m = 2) and read-most (m = 3). Recognition occurs when a consumer-specific threshold, $-\beta_{i,1,0}$ for ad-noted, $-\gamma_{i,2,0}$ for brand-associated, and $-\gamma_{i,3,0}$ for read-most, is exceeded (Hintzman 2000). This formulation extends the work of Wedel and Pieters (2000), who include fixation frequencies, rather than unobserved attention, in a binary probit memory model.

Because the original advertisement is available to the participants during the recognition test, we predict that the sizes of the three ad elements act as memory retrieval cues (Mitchell and Johnson 2000; Roediger and McDermott 2000). That is, consumers may use them to infer their prior attention to the advertisement and its elements. For example, a large pictorial element may lead consumers to infer that they must have seen the advertisement, and a large text element consisting of many words may lead consumers to believe that they probably read most of the text. We also predict that brand familiarity affects recognition memory because the fluency of processing the advertisement due to familiarity with the advertised brand may increase the likelihood of claiming ad recognition, independent of prior attention (Kelly and Jacoby 2000; Mitchel and Johnson 2000). Alternatively, familiarity may decrease the threshold for recognition because less attention may be required to store advertisements for familiar brands. We are not able to distinguish these two mechanisms of familiarity from our estimates. To allow for these effects, we include the size of the ad elements and brand familiarity in $x^M_{i,m,l}$ in Equation 4. The parameter vectors $\beta'_{i,1}$, $\gamma'_{i,2}$, and $\gamma'_{i,3}$ reflect the direct effects of these variables on the recognition measures, beyond their indirect effects mediated through attention to the advertisement or ad element.

Thus, the model we specify in Equations 1–4 enables us to test the ad layout's effects on attention, its indirect effects on ad recognition mediated by attention (MacKinnon, Fairchild, and Fritz 2007), and its direct effects on recognition beyond the effects through attention. These latter direct effects would demonstrate systematic biases in the recognition scores, which would reduce their diagnostic value.

Model Estimation

Because several of the (standard diffuse) prior distributions are not conjugate to the likelihood and the full conditional posteriors do not take on well-known forms, we use a Metropolis-within-Gibbs sampling algorithm to estimate the model (Rossi, Allenby, and McCulloch 2005). We use multivariate normal priors for all regression coefficients with mean zero and variance 10^4I. For the variance–covariance matrices D, we set inverse Wishart priors to have expectation I, with degrees of freedom equal to their rank plus one. We use 50,000 draws with a burn-in of 25,000, retaining every 50th target draw to reduce autocorrelation. Convergence is achieved well before the end of the burn-in. We tabulate posterior means and standard deviations. We compare the log-marginal densities (LMDs) of the proposed full model with those of simpler alternatives to gain insight into the contribution of each of the specific model components. To compute the LMDs, we use the methods proposed by Chib (1995) and Chib and Jeliazkov (2001) for the Gibbs sampler and Metropolis-within-Gibbs sampler. This involves a sequence of reduced Markov chain Monte Carlo (MCMC) runs for each of the models, in which sets of parameters are fixed at their posterior means, successively.

DIAGNOSTIC VALUE OF AD RECOGNITION

The proposed model predicts recognition memory from prior attention to the advertisements and other factors. Therefore, it can be used deductively to establish the proba-

[2]We have allowed the error terms of the model components m = 1, 2, 3 to covary, based on a suggestion by Peter Lenk (personal communication, 2007). These covariances were not significant in the application, and the model yielded similar estimates. We parameterized the heterogeneity distribution as a function of gender but did not find significant effects in the application, and thus we do not report these effects.

bility that recognition is claimed when consumers attended to the advertisement and its elements and the probability that recognition is not claimed when consumers did not attend to the advertisement and its elements (see Altman and Bland 1994a). In advertising research and practice, however, ad recognition tests are used inductively to make inferences about attention to advertisements during prior exposure in situations in which attention is not directly measured, for example, through eye tracking. In these applications, the researcher would like to know the accuracy of the recognition test as a diagnostic measure for attention. This inductive use is similar to the application of medical diagnostic tests (Altman and Bland 1994b; Guggenmoos-Holzmann and Van Houwelingen 2000). In that literature, the positive predictive value of a test is considered the proportion of people with positive test results who are accurately diagnosed to have the condition in question, and the negative predictive value is the proportion of people with negative test results who are accurately diagnosed not to have it (Altman and Bland 1994b; Phelps and Ghaemi 2006).

We propose to assess the diagnosticity of recognition tests through PDV and NDV, and we develop a procedure that provides BAR measures based on these metrics. We define PDV as the probability during exposure that a person has fixated on an advertisement or a specific ad element *at least* a certain number of times or more, given that he or she claims to have seen it. Similarly, we define NDV as the probability that a person has fixated on an advertisement or a specific ad element *at most* a certain number of times, given that he or she claims *not* to have seen it. Thus, these diagnostic values are the inverse conditional probabilities of fixating on an advertisement or an ad element conditional on claimed recognition of the advertisement or ad element. We can use Bayes' theorem to derive these predictive values (Goodman 1999).

For the PDV of the recognition test, we compute the conditional probability that consumer i fixates on element j of ad l more than a certain fixation threshold (χ_{PDV}) given claimed recognition. Similarly, we compute NDV as the probability that consumer i fixates on element j of ad l less than a certain threshold (χ_{NDV}), given no claimed recognition:

$$\text{PDV}(\chi) = p\left(n_{i,j,l} \geq \chi_{\text{PDV}} \middle| I_{m,l} = 1\right) \tag{5}$$

$$\text{NDV}(\chi) = p\left(n_{i,j,l} \geq \chi_{\text{NDV}} \middle| I_{m,l} = 0\right).$$

Equation 5 can be evaluated on the basis of the parameter estimates obtained from the attention and memory model using Bayes' theorem. That is, we compute the inverse probability that a person has fixated on the advertisement or ad element, given that he or she claims (no) recognition, $p(N_{i,j,l} = n_{i,j,l}|y_{i,m,l})$, as follows:

$$\left[\int_{\omega}\int_{\lambda}\int_{\mu}\int_{t} f_N\left(n_{i,j,l}\middle|\mu_{i,j,l}\right) f_{T|N}\left(t_{i,j,l}\middle|n_{i,t,l},\lambda_{i,j,l}\right) f_Y\left(y_{i,m,l}\middle|\omega_{i,m,l}\right) p\left(\mu_{i,j,l},\lambda_{i,j,l},\omega_{i,m,l}\middle|N,T,Y\right)\partial t\partial\mu\partial\lambda\partial\omega\right] \Bigg/ \left[\int_{\omega}\int_{\lambda}\int_{\mu} p\left(y_{i,m,l}\middle|\omega_{i,m,l}\right) p\left(\mu_{i,j,l},\lambda_{i,j,l},\omega_{i,m,l}\middle|N,T,Y\right)\partial\mu\partial\lambda\partial\omega\right]. \tag{6}$$

Here, $f_N(n_{i,j,l}|\mu_{i,j,l})$ is the conditional probability of observing $n_{i,j,l}$ fixations given $\mu_{i,j,l}$, and $f_{T|N}(n_{i,j,l}|\mu_{i,j,l}, \lambda_{i,j,l})$ is the conditional density of observing gaze duration $t_{i,j,l}$ given $n_{i,j,l}$ and $\lambda_{i,j,l}$ for ad element j associated with consumer i and ad l. In addition, $f_Y(y_{i,m,l} = 1|\omega_{i,m,l})$ is the probability that consumer i responds "yes" ("no" corresponds to $y_{i,m,l} = 0$) to recognition measure m (m = 1 for ad-noted, m = 2 for brand-associated, and m = 3 for read-most), given his or her latent attention $\phi_{i,l}$ to ad l and biases occurred in the memory process ($\omega_{i,m,l}$). Thus, $n_{i,j,l}$ is conditionally independent of $y_{i,m,l}$, given latent attention, $\phi_{i,l}$. Note that we use fixation frequency as the basis for computing the PDV and that the numerator in Equation 6 is integrated over $t_{i,j,l}$. Operationally, to compute the PDV and NDV for each of the three recognition measures (ad-noted, brand-associated, and read-most) for each advertisement, in the MCMC chain after the burn-in period, we first compute $f_N(n_{i,j,l} < \chi|\mu_{i,j,l})$ and $f_Y(y_{i,m,l} = 1|\omega_{i,m,l})$ on the basis of Equations 1 and 3, respectively. The term $f_Y(y_{i,m,l} = 1|\omega_{i,m,l})$ is used for the denominator of the PDV, and $1 - f_Y(y_{i,m,l} = 1|\omega_{i,m,l})$ is used for the denominator of the NDV. Next, we compute $[1 - f_N(n_{i,j,l} < \chi|\mu_{i,j,l})] \times f_Y(y_{i,m,l} = 1|\omega_{i,m,l})$ for the numerator of the PDV and $f_N(n_{i,j,l} < \chi|\mu_{i,j,l}) \times [1 - f_Y(y_{i,m,l} = 1|\omega_{i,m,l})]$ for the numerator of the NDV. After running the MCMC, we average numerator draws and then denominator draws for each advertisement to integrate out $t_{i,j,l}$, $\mu_{i,j,l}$, $\lambda_{i,j,l}$, and $\omega_{i,m,l}$ to compute the PDV and NDV.

The higher the value of the PDV metric for a specific threshold χ_{PDV}, the more diagnostic the recognition measure is for prior attention to the advertisement or its elements. The higher the value of the NDV metric for a specific threshold χ_{NDV}, the more diagnostic the recognition measure is for no prior attention to the advertisement or its elements. Because these metrics are derived as an integral part of the model that accounts for the influence of explanatory variables, they are independent of these explanatory variables and unbiased, as desired for diagnostic tests (Leisenring and Pepe 1998).

We derive diagnostic values of ad recognition measures as part of the MCMC runs using Bayes' theorem, which is preferable to previously used plug-in estimators (Rossi, Allenby, and McCulloch 2005), and we demonstrate how these diagnostic values can be used in a bias-adjustment procedure for ad recognition measures.

RESULTS

We compare the LMD of several nested alternative models to determine the contribution of specific factors to recognition memory, with a higher LMD indicating stronger support for the model in question. We begin with a baseline model that contains only the effects of ad layout and brand familiarity on attention and the effects of attention on recognition memory. It rests on the assumption that ad layout and brand familiarity effects on recognition are completely mediated by attention (Zhang, Wedel, and Pieters 2009). Support for the model would imply that the recognition measures are unbiased in reflecting attention during prior ad exposure. The LMD of the baseline model is –79,816. The second model, which adds the direct effects of the brand, pictorial, and text size on ad recognition, improves on this (LMD increases to –79,495). Thus, ad layout directly influences ad recognition, beyond its effects mediated by atten-

tion. The third model, which adds the direct effects of brand familiarity on ad recognition to Model 2, further improves on this (LMD increases to –79,346). Thus, brand familiarity directly influences ad recognition, beyond its effects mediated by attention. Collectively, these findings reveal that the ad recognition measures do not purely reflect attention to prior ad exposure but are indeed biased as a result of memory retrieval factors. We present parameter estimates of the third model.[3]

Table 2 presents the parameter estimates for the attention part of the model. In line with previous research (Pieters and Wedel 2004), the effect of size of the text element on fixation frequency on the text is the largest, followed by that of the size of the brand on its fixation frequency, and finally by that of the pictorial on fixation frequency on the pictorial. The large effect of the size of the text element is most likely due to the more focal, serial processes during reading (Reichle, Rayner, and Pollatsek 2003), whereas the gist of pictorials can often be grasped in a glance (Rayner 1998). In general, ad elements compete for attention, as shown by the significant, negative cross-effects of their sizes; for example, larger pictorial sizes reduce attention to the brand. There is a positive cross-effect of brand size on attention to the pictorial, which may capture a positive transfer of brand information to pictorial attention (Pieters and Wedel 2004). More familiar brands receive higher fixation frequencies to the pictorial and the text. This shows that, consistent with prior research, ad layout and brand familiarity influence attention to advertisements.

Table 3 presents the parameter estimates for the recognition part of the model. There is clear evidence for attention effects on the ad-noted measure and for the brand attention effect on the brand-associated measure. This supports the validity of these recognition measures as indicators of ad attention. However, the read-most measure is not significantly affected by attention to the text of advertisements. Table 3 also shows that ad layout has direct effects on recognition memory, beyond those mediated by attention. A

[3]We also estimated a version of the memory model in which we include observed fixations instead of latent attention and find that the estimates of the two models are comparable. The full model is preferable on theoretical grounds.

Table 2
DETERMINANTS OF AD ATTENTION

	Attention to Advertising					
	Brand		*Pictorial*		*Text*	
Predictors	M	SD	M	SD	M	SD
Fixation Frequency						
Intercept	**.769**	.048	**1.640**	.035	**.896**	.054
Surface Size						
Brand	**1.231**	.050	**.128**	.039	**–.785**	.057
Pictorial	**–.611**	.055	**.847**	.049	**–.095**	.056
Text	**–.557**	.050	**–.215**	.037	**1.742**	.045
Brand Familiarity	.023	.028	**.049**	.025	**.098**	.030
Covariances for Fixation Frequency						
Brand	**.403**	.043	(.583)		(.737)	
Pictorial	**.176**	.027	**.227**	.024	(.543)	
Text	**.336**	.45	**.186**	.033	**.519**	.186
Fixation Duration						
Ln(Mean)	**–1.573**	.017	**–1.639**	.016	**–1.596**	.017

Notes: Bold parameter estimates indicate that probabilities of the parameters to be larger or smaller than zero are greater than .95. Correlations are in parentheses.

Table 3
DETERMINANTS OF AD RECOGNITION

	Ad Recognition Memory					
	Ad Noted		*Brand Associated*		*Read Most*	
Predictors	M	SD	M	SD	M	SD
Intercept	**.769**	.048	**1.640**	.035	**.896**	.054
Latent Attention						
Brand	.128	.137	**.337**	.120	.068	.120
Pictorial	**.447**	.102	–.080	.091	.018	.094
Text	**.264**	.075	**–.128**	.062	–.026	.061
Surface Size						
Brand	**–.345**	.125	**–.276**	.123	**–.609**	.134
Pictorial	**.278**	.126	.074	.144	.138	.155
Text	.036	.107	.122	.114	**.342**	.116
Brand Familiarity	**.137**	.032	–.011	.030	**.067**	.032

Notes: Bold parameter estimates indicate that probabilities of the parameters to be larger or smaller than zero are greater than .95. Correlation between brand-associated and read-most measures is .158 (SD = .020).

larger pictorial increases ad-noted, regardless of how much attention was devoted to the advertisement during the prior exposure. These findings are consistent with findings on the effect of pictorial size on ad recognition measures (Finn 1988), but we show that the effect is independent of the actual attention devoted to the pictorial. Thus, larger pictorials in advertisements lead to a systematic overclaiming of prior attention to the advertisements.

In addition, more text in the advertisement increases the probability of claiming recognition of text, regardless of how much attention was actually paid to the elements. The large positive direct effect of text size on the read-most measure is particularly troublesome because, though text size influences attention to text, attention to text does not subsequently influence text recognition. Conversely, larger brand sizes decrease the ad-noted, brand-associated, and read-most measures, independent of the actual attention devoted to them during ad exposure.[4] Apparently, larger text and smaller brand elements serve as retrieval cues at the time of the recognition test, which lead people to infer that more attention must have been devoted to the text during ad exposure. In addition, people claim to have noted advertisements for familiar brands more often and to have read most of their text, independent of their actual attention to the advertisements. This finding, along with the finding that familiar advertisements receive more fixations on the pictorial and text but not the brand, indicates that familiarity with the brand may lower the threshold for ad recognition.

Taken together, these results reveal that whereas recognition memory for the advertisement as a whole and its brand element reflects prior attention to some extent, memory for text is mostly reconstructed during the recognition test and bears little relation to attention at exposure. Moreover, all measures of recognition memory are systematically influenced by factors other than actual attention during ad exposure, which shows that they are biased.

BIAS-ADJUSTMENT OF RECOGNITION MEASURES

Figure 1 provides the positive and negative diagnosticity curves. The curves plot the PDV and NDV as computed from the parameter estimates for the ad-noted, brand-associated, and read-most measures, averaged across advertisements and consumers, against values of the fixation threshold (χ = 0, 1, 2, …). Figure 1 also depicts the interval containing 90% of the advertisements, for each of these curves.[5] In interpreting the PDV and NDV and debiasing the recognition scores, we focus on $\chi_{PDV} = \chi_{NDV} = 5$. Although other thresholds are readily accommodated, five fixations are a natural cutoff in eye-tracking studies of complex scenes such as advertisements and have been used in a range of studies by, among others, Charness and colleagues (2001), Masciocchi and colleagues (2008), and Torralba and colleagues (2006). This threshold corresponds to approximately 1–2 seconds of exposure needed for reliable recognition memory, which reflects exposure durations to advertisements in natural conditions for the majority of people (Pieters and Wedel 2004). We tried different values of the thresholds, and the results are fairly stable across a small range of values (four to six) around the five-fixation threshold but may change when substantially larger or smaller thresholds are chosen. Therefore, we believe that $\chi_{PDV} = \chi_{NDV} = 5$ is a reasonable choice in many studies. We investigate its validity further.

The top panel of Figure 1 shows that the ad-noted measure has the highest PDV. If consumers claim to recognize the advertisement (PDV ad-noted), the probability of having had, on average, five or more fixations is 92.6%. For the brand-associated and read-most measures, the probabilities of having had, on average, five or more fixations, given claimed recognition, are much lower: 26.2% and 35.8%, respectively. To illustrate, at the threshold, the odds of the PDVs of ad-noted over brand-associated are almost 4:1 (.93/.26) in favor of ad-noted. However, the number of fixations on the brand and text is smaller than that on the advertisement as a whole (Table 1).

The bottom panel of Figure 1 shows that the brand-associated measure has the highest NDV. If consumers claim not to have noted the brand in the advertisement (NDV brand-associated), the probability of having had fewer than five fixations during the original ad exposure is 81.6%. If they claim not to have read most of the text (NDV read-most), the probability of having had fewer than five fixations is 71.8%, which is also fairly high. However, if consumers claim not to have noted the advertisement (NDV ad-noted), the probability of having had fewer than five fixations is only 12.2%. This suggests that claims not to have noted the advertisements are unreliable and that false negative claims are common as long as consumers fixate on an advertisement for less than 1–2 seconds. To illustrate, at the fixation threshold, the odds of brand-associated over ad-noted NDV are close to 7:1 (.82/.12) in favor of brand-associated.

Thus, the ad-noted measure has the highest PDV but, at the same time, the lowest NDV, while the reverse holds for the brand-associated measure. If consumers claim to have noted an advertisement, there is a high probability (92.6%) that they fixated on the advertisement at least five times, and if they claim not to have noted the brand in the advertisement, there is fairly high probability (81.6%) that they fixated on it fewer than five times.

On the basis of this finding, we propose to use the PDV and NDV as bias-adjustment factors for the ad recognition measures. That is, raw recognition scores indicate the proportion of consumers who claim to recognize an advertisement and its elements, even when they may not actually have attended them. When these raw recognition scores are available but eye-tracking data are not, it may be useful to adjust the raw scores to remove biases. The BAR scores indicate the estimated proportion of consumers who fixated on the advertisement or its elements five times or more. Our proposed adjustment uses values of PDV(χ) and NDV(χ) that can be read directly from Figure 1 at the threshold χ = 5 (or any other desired threshold). We can compute BAR scores as follows:

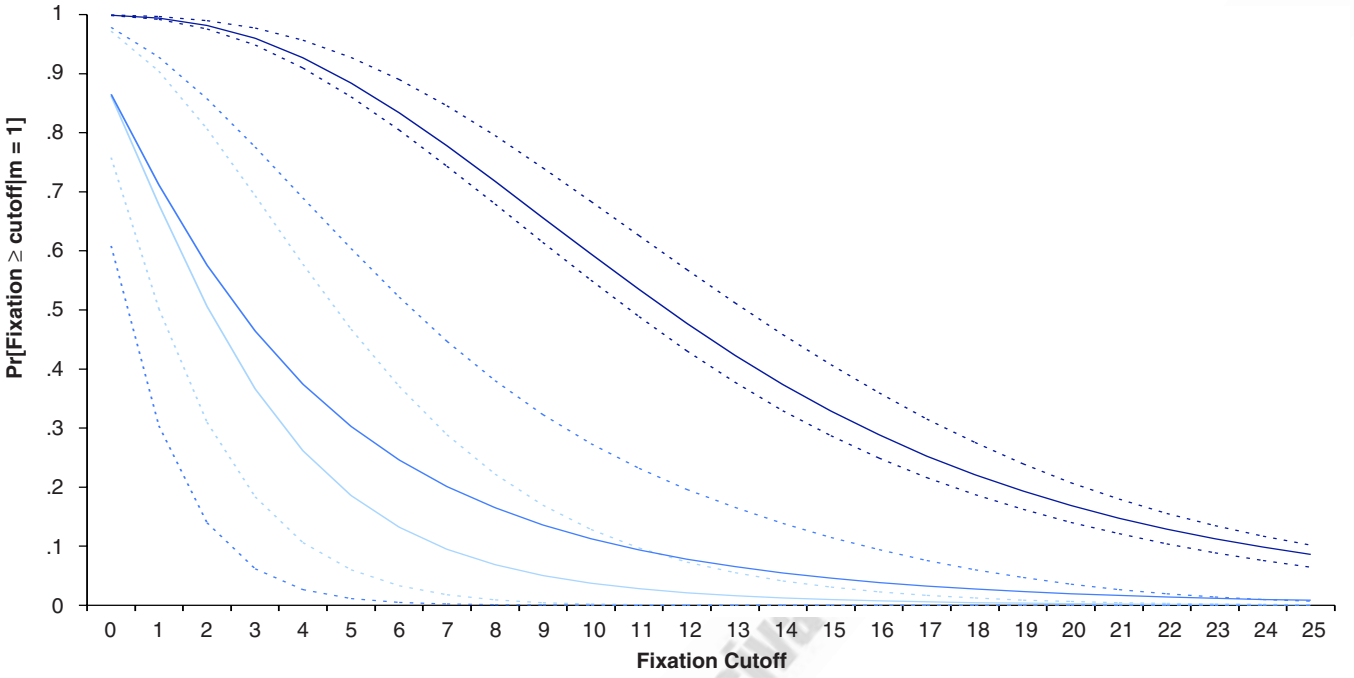
$$\text{BAR score} = \text{PDV}(\chi) \times \{\text{Raw score}\} + [1 - \text{NDV}(\chi)] \times \{1 - \text{Raw score}\}. \quad (7)$$

[4]This negative effect is not due to collinearity with attention or the other surface sizes, because it persists even when these other variables are eliminated from the model.

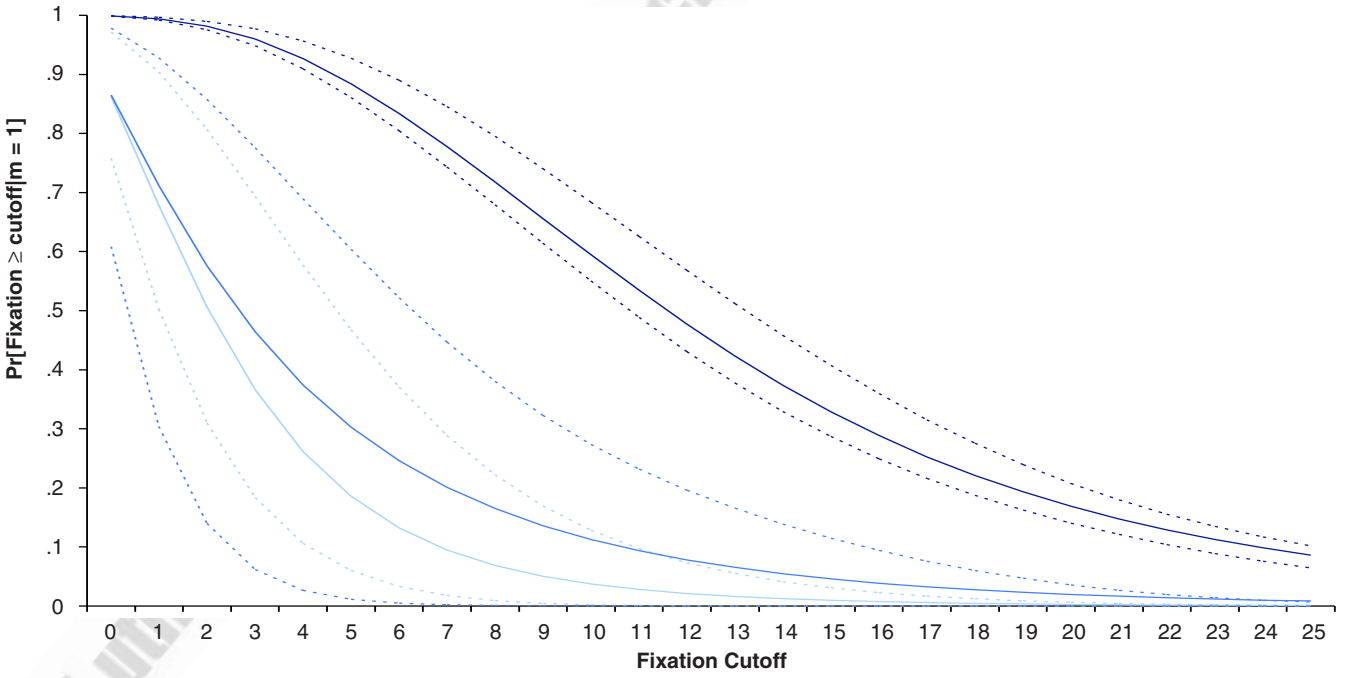
[5]The Web Appendix (http://www.marketingpower.com/jmrjune10) provides the diagnostic value curves separately for each of the 48 advertisements in this study.

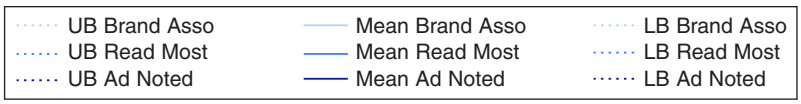
Figure 1
POSITIVE AND NEGATIVE DIAGNOSTICITY CURVES

A: Positive Diagnostic Curve



B: Negative Diagnostic Curve





Notes: PDV [pr(number fixation ≥ fixation cutoff | m = 1)] and NDV [pr(number fixation < fixation cutoff | m = 0)] at different fixation thresholds.

Equation 7 is derived from the rule of total probability: $P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$. Here, P(A) is the quantity required but unknown from a recognition test: the probability that consumers fixate on an advertisement (or the brand or text element) five times or more (the BAR score). As such, P(A|B) is the probability that consumers fixate on the advertisement five times or more, given claimed recognition, which is the PDV; P(B) is the probability that consumers claim ad recognition (raw recognition score); $P(A|\bar{B})$ is the probability that consumers fixate on the advertisement five times or more, given no claimed recognition, which equals (1 – NDV); and $P(\bar{B})$ is the probability that consumers do not claim ad recognition (1 – raw recognition score from the test).

In this way, the BAR score provides information about attention during ad exposure given claimed ad recognition. Importantly, the BAR score can be computed using Equation 7 for new samples of advertisements and consumers for which only the recognition scores, but not the eye-tracking measures, are available. For example, if the raw ad-noted score is .80, and the PDV and NDV given the threshold ($\chi_{PDV} = \chi_{NDV} = 5$) are computed to be .93 and .12, respectively, the BAR score is (.93)(.80) + (.88)(.20) = .92. In general, the BAR score can range from 0 to 1. When PDV and NDV approximately sum to one, the bias-adjusted test approximately equals the PDV. All else being equal, the BAR score increases as the PDV increases and decreases as the NDV increases. The final adjustment depends on the balance between these two.

To investigate the diagnostic values further, we regressed the log-odds diagnosticity [log(PDV/(1 – NDV)] for each advertisement on its associated brand popularity, ad attractiveness, and ad uniqueness ratings, as well as the number of words in the text for each of the three recognition measures (Table 4). The higher the log-odds diagnosticity, the more diagnostic the recognition measure in question is for prior attention during exposure. The diagnostic value of the ad-noted score is higher for advertisements that are unique and have more words of text in the headline. The diagnostic value of the brand-associated measure increases with ad attractiveness but decreases with the amount of text in the headline. Finally, brand popularity has a negative effect while ad attractiveness has a positive effect on the diagnosticity of the read-most measure.

Holdout Validation

To demonstrate the improved accuracy of BAR scores over raw recognition scores, we use two holdout samples. First, we reestimate the model for all participants and a random sample of 38 advertisements and retain 10 advertisements as a holdout sample. Second, we reestimate the model for a random sample of 38 advertisements and 145 participants and retain 10 advertisements and 40 participants as a holdout sample. The first holdout sample enables us to assess the performance of our approach for a sample of new advertisements for the same participants in the test, and the second holdout sample enables us to assess performance for a new sample of advertisements and a new sample of participants. We adjust the raw scores of the holdout sample of advertisements using Equation 7, with PDV and NDV estimated from the calibration sample, averaging PDV and NDV across participants and advertisements for each of the recognition measures. We define the true score as the proportion of consumers who actually fixated on the advertisement or the brand and text elements five or more times and compute the absolute deviations of these BAR scores from the true scores |BAR score – true score| and of the raw scores from the true scores |raw score – true score| for each advertisement. Averaging these absolute deviations across the advertisements in the holdout sample, we obtain the mean absolute deviations (MAD) of the raw (MAD_r) and BAR (MAD_b) scores. Table 4 provides the in-sample and out-of-sample results.

As we expected, BAR scores are more accurate than the raw scores in reflecting actual fixations on the advertisement and its elements, both in-sample and out-of-sample, for all three recognition measures (i.e., all $MAD_b < MAD_r$). The results for the sample of new advertisements/same participants and of new participants/new advertisements are similar, so we only discuss the latter in detail. In-sample MADs of the BAR scores are relatively small: 10.2%, 13.7%, and 18.2%, respectively, for ad-noted, brand-associated, and read-most. These are large reductions from the MADs for the raw scores, which are approximately 25% (Table 4). Not surprisingly, the BAR score for read-most still performs worst. This is due to the absence of a significant relationship between text attention and recognition, which gives the bias-adjustment procedure little to work with. Whereas the out-of-sample MAD (9.8%) for the

Table 4
BIAS ADJUSTMENT OF RECOGNITION MEMORY

	Ad Recognition Memory		
	Ad Noted	*Brand Associated*	*Read Most*
PDV (%)	92.6	26.2	35.8
NDV (%)	12.2	81.6	71.8
ln(PDV) – ln(1 – NDV)	.395	.302	.054
Regression Analysis			
Intercept	**.287**	**.262**	**.053**
Brand popularity	.001	.012	**–.002**
Ad uniqueness	***.039***	**–.012**	–.001
Ad attractiveness	–.026	***.042***	**.003**
Number of words	**.015**	**–.018**	.000
In-Sample (Ad only in %)			
MAD raw score (MAD_r)	27.8	24.1	24.1
MAD bias-adjusted score (MAD_b)	9.7	14.2	18.4
% improvement in MAD	65.1	41.1	23.7
Out-of-Sample (Ad only in %)			
MAD raw score (MAD_r)	21.8	23.6	21.8
MAD bias-adjusted score (MAD_b)	9.8	8.9	12.9
% improvement in MAD	55.1	62.3	40.8
In-Sample (Both Ad and Participant in %)			
MAD raw score (MAD_r)	26.8	24.4	23.8
MAD bias-adjusted score (MAD_b)	10.2	13.7	18.2
% improvement in MAD	61.9	43.9	23.5
Out-of-Sample (Both Ad and Participant in %)			
MAD raw score (MAD_r)	23.5	25.0	19.0
MAD bias-adjusted score (MAD_b)	9.8	10.1	14.0
% improvement in MAD	58.3	59.6	25.8

Notes: % improvement in MAD = $(MAD_r - MAD_b)/MAD_r \times 100$. Bold (bold and italic) parameter estimates indicate statistical significance at $\alpha = .05$ (.10).

ad-noted score is close to the in-sample MAD, the out-of-sample MADs are even somewhat smaller for the brand-associated (10.1%) and read-most (14.0%) scores. This may be due to the specific advertisements in our (random) hold-out sample. The magnitudes of these out-of-sample MADs (new advertisements/new participants) indicate good performance of the bias adjustment procedure. We also compute the percentage improvement in BAR scores relative to MAD_r (Table 4). Improvement in accuracy ranges from approximately 25% to 60% out-of-sample, which is substantial.[6]

Bias-Adjustment for Ad Recognition in the In-Home Setting

So far, the reported results were obtained from data collected in a laboratory setting, which enabled us to collect both eye movements and recognition measures from the same people. Yet bias adjustment seems particularly valuable when ad recognition testing takes place under natural exposure conditions, in which attention to advertisements is short (approximately a few seconds; see Pieters and Wedel 2004) and eye-tracking measures are typically unavailable. Therefore, we also apply the bias-adjustment procedure to the data collected after in-home exposure. Eye-tracking data are not available for the participants in the in-home condition. To explore the effects of bias adjustment in this setting, we compare the recognition scores between the in-home and the lab conditions before and after bias adjustment. We randomly allocated participants to one of the two conditions, and they evaluated the same set of advertisements in the same editorial context. If indeed common retrieval biases are removed with the bias-adjustment procedure, the scores of the in-home and the lab conditions should be closer after our correction.

In Table 5, the differences between the raw recognition scores in the in-home and the lab conditions are substantial: 16.0% for ad-noted, 12.2% for brand-associated, and 6.7% for read-most. For the lab condition, we again observe that bias-adjusted scores are closer than the raw scores to true scores. After bias correction, however, the differences between the in-home and the lab conditions diminish substantially to .7%, .6%, and .5%, respectively. These results are due in part to the low diagnosticity of the test, but they reveal that bias adjustment reduces the gap between the recognition scores of the in-home and the lab conditions and corrects recognition scores collected after exposure in natural in-home settings, as is frequently done in practice. This supports the potential improvements due to the proposed bias-adjustment procedure.

[6]We also corrected Starch scores using PDV and NDV computed simply as the proportion of participants who fixated on an advertisement greater than or equal to (less than) the fixation threshold given that they reported having (not) seen the advertisement. This correction leads to a worse prediction than raw Starch scores.

Table 5
BIAS ADJUSTMENT FOR LAB VERSUS IN-HOME DATA

	Ad Recognition Memory		
	Ad Noted (%)	*Brand Associated* (%)	*Read Most* (%)
True score at the fixation threshold (lab data)	80.5	20.6	29.8
Raw Score			
Lab	54.7	41.1	17.2
In-home	39.1	29.6	16.9
MAD between lab and in-home	16.0	12.2	6.7
Bias-Adjusted Score			
Lab	90.4	21.3	30.0
In-home	89.7	20.5	29.9
MAD between lab and in-home	.7	.9	.5

DISCUSSION

Diagnosticity of Ad Recognition Scores

We found that attention to the ad predicted the ad-noted measure and that attention to the brand predicted the brand-associated measure. This is good news because it demonstrates a certain diagnostic value of these ad recognition measures. However, attention to the text in advertisements did not significantly affect the read-most measure. Independent of attention, consumers overclaimed ad recognition when the advertisement contained a larger pictorial and a smaller-sized brand (ad-noted and read-most) and when the text portion was larger (brand-associated and read-most). This configuration of larger pictorials, smaller brands, and larger text represents a typical ad layout. Thus, regardless of actual attention to them during prior exposure, recognition of advertisements with prototypical layouts was overclaimed in recognition tests. Failure to control for ad prototypicality in ad recognition measures may lead to overrating of the effectiveness of the specific advertisements.

The diagnostic value of ad recognition is low but varies across measures and metrics. Specifically, the PDV of the ad-noted measure was high, but its NDV was low, so the ad-noted measure was best at identifying advertisements that were actually noted. Conversely, the brand-associated and read-most scores had lower PDVs but higher NDVs, so the brand-associated and read-most recognition measures were better at excluding advertisements for which the brand element was actually not identified and the text was not read. None of the ad recognition measures performed well in both accurately identifying attended and excluding unattended advertisements. Thus, a worthwhile question for further research is whether the diagnostic value of some representational forms of the brand element, including logo, brand name, and brand slogan, is better, which would necessitate the collection of fixation and recognition data for such representations separately.

BAR Scores

Starch-type recognition measures have a long tradition in advertising practice and are relatively easy and cheap to collect. Although we believe that eye-tracking measures are superior measures of attention, discarding recognition measures may lead to undesirable regime switches in measurement of ad effectiveness for a large number of companies relying on them. Therefore, research is called for to improve the accuracy of measurement instruments for ad recognition tests. This could be done, for example, by asking test participants to provide confidence judgments using "consider the opposite" strategies or cueing debiasing factors (Arkes 1991). Triangulation with other memory meas-

ures, such as recall and indirect measures of memory, is another viable route (Krishnan and Chakravarti 1999).

We proposed BAR scores that indicate the proportion of consumers who fixated on the advertisement or its elements five times or more. The proposed BAR scores may be gainfully used to remove biases from recognition scores when practical considerations dictate the continued use of these recognition scores and eye-tracking measures are not available. In these cases, the bias adjustments may improve the accuracy of recognition memory scores for 1–2 seconds' exposure durations by as much as 25%–60% and remove some of the differences between tests conducted after in-home and lab exposure conditions. However, to assess attention to print advertisements, we believe that eye-tracking measures, if available, are preferable to adjusted recognition measures to assess attention to advertisements.

Because we could not track consumers' eye movements at home, we were not able to assess biases for that condition directly. However, because the memory traces in the in-home condition were even weaker than in the lab condition, recognition memory may have been even more biased than what we observed in the lab condition. Although the results on the reduction of biases in the in-home condition may be consistent with the presence of common retrieval biases, other, more mechanical explanations cannot be excluded. Further research might address these issues.

We chose the threshold of five fixations, which we used for each of the three ad recognition measures, on the basis of theory and prior research. However, other choices are possible, and different thresholds could be used for different recognition measures. We conducted a sensitivity analyses of diagnosticity to threshold values (Altman and Bland 1994c), which showed that the total diagnosticity (i.e., sum of PDV and NDV, with two as the theoretical maximum) never exceeded 1.14 for any threshold value, which is low. For the brand-associated measure, there was no threshold in which the PDV and NDV both exceeded .50. The sensitivity analyses showed that the bias adjustments are fairly robust across a small range of thresholds around the five-fixation threshold, and we conclude that the threshold of five fixations is a reasonable one. Although the findings cast doubts on the diagnostic value of ad recognition measures for attention during prior ad exposure, which they purport to reflect, it is not clear below which specific PDV and NDV ad recognition tests are still useful. This, along with the optimal choice of the recognition thresholds, is an important topic for further research.

Implications for Theory and Practice

In academic advertising research, the use of ad recognition measures may misdirect theory development. In the current study, for example, larger pictorials increased the ad-noted measure substantially, independent of the actual attention devoted to the advertisement. This may lead to overvaluing of the role of the pictorial at the expense of the text and brand in determining attention to advertising (Finn 1988, 1992; Mothersbaugh, Huhmann, and Franke 2002). More generally, using recognition memory to infer the influence of stimulus and person factors on attention during ad exposure and/or on recognition during memory retrieval is tricky (Puntoni and Tavassoli 2007; Whittlesea and Leboe 2000) because such factors may influence both exposure and retrieval and in quite different ways. For example, in this study, the size of the text element decreased attention to the brand but increased brand recognition, independent of attention. Without measures of attention during ad exposure, only effects on memory remain without insights into how they arise.

In advertising practice, ad recognition measures are used in pre- and posttesting and campaign evaluation, with some practitioners even calling them "the definitive advertising measurement scores."[7] Thus, ad-noted, brand-associated, and read-most scores across magazines and product categories have been used to benchmark the effectiveness of print advertising,[8] and similar recognition measures are used in television advertising.[9] They are being used to assess which advertisements attract the most attention, and they serve as inputs to advertising message and media decisions. Our findings raise doubts about the validity of the current ad recognition measures for these purposes: Such measures are not strong proxies for attention, and in particular, text recognition is not related to attention at all. Memory biases may especially harm prototypical advertisements because their ad-noted and read-most scores tend to be overvalued, independent of actual attention during exposure. Comforted by high recognition scores, advertisements may then be insufficiently optimized, and their campaigns may be sustained beyond the cost-effective level of repeated exposures. Benchmarking advertisements against other advertisements on the basis of raw ad recognition measures requires caution, given the wide variations in PDV and NDV across advertisements (see the Web Appendix at http://www.marketingpower.com/jmrjune10). A reason that ad recognition measures are recommended in advertising research is their presumed ability to detect delicate attentional and perceptual processes during exposure (Heath and Nairn 2005). The current findings indicate that these measures may unfortunately have insufficient diagnostic value for this purpose.

Although this research focused on diagnosticity of Starch recognition tests for print advertisements, the proposed framework can be useful in other tests situations in marketing research as well, such as recognition tests of outdoor, television, and Web advertising and unaided/aided recall scores. Only after advertisements have been diagnosed accurately for their prior exposure can attempts to improve their future performance become effective. We hope that the proposed framework for diagnosticity and bias adjustment of recognition tests contributes to improved performance by "raising the BAR."

REFERENCES

Allenby, Greg M. and Peter Rossi (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89 (1–2), 57–78.

Altman, Douglas G. and J. Martin Bland (1994a), "Diagnostic Tests 1: Sensitivity and Specificity," *British Medical Journal*, 308 (6943), 1552.

[7]See http://www.mcnairingenuity.com.au (accessed July 2008).

[8]See http://findarticles.com/p/articles/mi_m4PRN/is_2008_June_3/ai_n25475031 and http://www.gfkamerica.com/practice_areas/brand_and_comm/starch/adnorms/index.en.html (accessed July 2008).

[9]See http://www.ameritest.net/products/adtracking.pdf (accessed July 2008).

——— and ——— (1994b), "Diagnostic Tests 2: Predictive Values," *British Medical Journal*, 309 (6947), 102.
——— and ——— (1994c), "Diagnostic Tests 3: Receiver Operating Characteristic Plots," *British Medical Journal*, 309 (6948), 188.
Arkes, Hal R. (1991), "Costs and Benefits of Judgment Errors: Implications for Debiasing," *Psychological Bulletin*, 110 (3), 486–98.
Bagozzi, Richard P. and A.J. Silk (1983), "Recall, Recognition, and the Measurement of Memory for Print Advertisements," *Marketing Science*, 2 (2), 95–134.
Baldinger, Allan L. and William A. Cook (2006), "Ad Testing," in *Handbook of Marketing Research*, Rajeev Grover and Marco Vriens, eds. London: Sage Publications, 487–505.
Belch, George E. and Michael A. Belch (2001), *Advertising and Promotion: An Integrated Marketing Communications Perspective*, 5th ed. Boston: McGraw-Hill.
Bhargava, Mukesh, Naveen Donthu, and Rosanne Caron (1994), "Improving the Effectiveness of Outdoor Advertising," *Journal of Advertising Research*, 34 (2), 46–55.
Charness, Neil, Eyal M. Reingold, Mark Pomplun, and Dave M. Stampe (2001), "The Perceptual Aspect of Skilled Performance in Chess: Evidence from Eye Movements," *Memory and Cognition*, 29 (8), 1146–52.
Chib, Siddhartha (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90 (432), 1313–21.
——— and Ivan Jeliazkov (2001), "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96 (453), 270–81.
——— and R. Winkelmann (2001), "Markov Chain Monte Carlo Analysis of Correlated Count Data," *Journal of Business & Economic Statistics*, 19 (4), 428–35.
Duchowski, Andrew T. (2003), *Eye Tracking Methodology: Theory and Practice*. London: Springer-Verlag.
Edwards, Yancy D. and Greg M. Allenby (2003), "Multivariate Analysis of Multiple Response Data," *Journal of Marketing Research*, 40 (August), 321–34.
Finn, Adam (1988), "Print Ad Recognition Readership Scores: An Information Processing Perspective," *Journal of Marketing Research*, 25 (May), 168–77.
——— (1992), "Recall, Recognition, and the Measurement of Memory for Print Advertisements: A Reassessment," *Marketing Science*, 11 (1), 95–100.
Goodman, Steven N. (1999), "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor," *Annals of Internal Medicine*, 130 (12), 1005–1013.
Guggenmoos-Holzmann, Irene and Hans C. van Houwelingen (2000), "The (In)Validity of Sensitivity and Specificity," *Statistics in Medicine*, 19 (1), 1783–92.
Hanssens, Dominique M. and Barton A. Weitz (1980), "The Effectiveness of Industrial Print Advertisements Across Product Categories," *Journal of Marketing Research*, 17 (August), 294–306.
Harris, Christopher, Louise Hainline, Israel Abramov, Elizabeth Lemerise, and Cheryl Camenzuli (1988), "The Distribution of Fixation Durations in Infants and Naive Adults," *Vision Research*, 28 (3), 419–32.
Havlena, William J. and Jeffrey Graham (2004), "Decay Effects in Online Advertising: Quantifying the Impact of Time Since Last Exposure on Branding Effectiveness," *Journal of Advertising Research*, 44 (4), 327–32.
Heath, Robert and Agnes Nairn (2005), "Measuring Affective Advertising: Implications of Low Attention Processing on Recall," *Journal of Advertising Research*, 45 (2), 269–81.
Heller Gillian Z., D. Mikis Stasinopoulos, Robert A. Rigby, and Piet de Jong (2007), "Mean and Dispersion Modelling for Policy Claims Costs," *Scandinavian Actuarial Journal*, 107 (4), 281–92.
Henderson, John M. (1992), "Object Identification in Context: The Visual Processing of Natural Scenes," *Canadian Journal of Psychology*, 46 (3), 319–41.
Hermie, Patrick, Trui Lankcriet, Koen Lansloot, and Stef Peeters (2005), *StopWatch. Everything of the Impact of Advertisements in Magazines*, (January), (accessed March 2009), [available at http://www.ppamarketing.net/cgi-bin/go.pl/research/article.html?uid=116].
Hintzman, Douglas L. (2000), "Memory Judgments," in *The Oxford Handbook of Memory*, Endel Tulving and Fergus I.M. Craik, eds. Oxford: Oxford University Press, 165–78.
Itti, Laurent (2005), "Models of Bottom-Up Attention and Saliency" in *Neurobiology of Attention*, Laurent Itti, Geraint Rees, and John K. Tsotsos, eds. Amsterdam: Elsevier Academic Press, 576–82.
Janiszewski, Chris (1998), "The Influence of Display Characteristics on Visual Exploratory Search Behavior," *Journal of Consumer Research*, 25 (December), 290–301.
Johnson, N.L., S. Kotz, and N. Balakrishnan (1994), *Continuous Univariate Distributions*, Vols. 1 and 2. New York: John Wiley & Sons.
Kelly, Colleen M. and Larry L. Jacoby (2000), "Recollection and Familiarity," in *The Oxford Handbook of Memory*, Endel Tulving and Fergus I.M. Craik, eds. Oxford: Oxford University Press, 215–28.
Krishnan Shanker H. and Dipankar Chakravarti (1999), "Memory Measures for Pretesting Advertisements: An Integrative Conceptual Framework and a Diagnostic Template," *Journal of Consumer Psychology*, 8 (1), 1–37.
Leisenring, Wendy and Margaret Sullivan Pepe (1998), "Regression Modelling of Diagnostic Likelihood Ratios for the Evaluation of Medical Diagnostic Tests," *Biometrics*, 54 (2), 444–52.
MacKinnon, David P., Amanda J. Fairchild, and Matthew S. Fritz (2007), "Mediation Analysis," *Annual Review of Psychology*, 58 (January), 593–614.
Manchanda, Puneet, Asim Ansari, and Sunil Gupta (1999), "The 'Shopping Basket': A Model for Multicategory Purchase Incidence Decisions," *Marketing Science*, 18 (2), 95–114.
Masciocchi, Christopher, Stefan Mihalas, Derrick Parkhurst, and Ernst Niebur (2008), "Interesting Locations in Natural Scenes Draw Eye Movements," *Journal of Vision*, 8 (6), 114.
Mehta, Abhilasha and Scott C. Purvis (2006), "Reconsidering Recall and Emotion in Advertising," *Journal of Advertising Research*, 46 (1), 49–56.
Mitchell, Karen J. and Marcia K. Johnson (2000), "Source Monitoring," in *The Oxford Handbook of Memory*, Endel Tulving and Fergus I.M. Craik, eds. Oxford: Oxford University Press, 179–95.
Mothersbaugh, David L., Bruce A. Huhmann, and George R. Franke (2002), "Combinatory and Separative Effects of Rhetorical Figures on Consumers' Effort and Focus in Ad Processing," *Journal of Consumer Research*, 28 (4), 589–602.
Phelps, James R. and S. Nassir Ghaemi (2006), "Improving the Diagnosis of Bipolar Disorder: Predictive Value of Screening Tests," *Journal of Affective Disorders*, 92 (2), 141–48.
Pieters, Rik and Michel Wedel (2004), "Attention Capture and Transfer in Advertising: Brand, Pictorial, and Text-Size Effects," *Journal of Marketing*, 68 (April), 36–50.
——— and ——— (2007), "Informativeness of Eye Movements for Visual Marketing: Six Cornerstones," in *Visual Marketing: From Attention to Action*, Michel Wedel and Rik Pieters, eds. New York: Lawrence Erlbaum Associates/Taylor & Francis, 43–71.
Puntoni, Stefano and Nader T. Tavassoli (2007), "The Effect of Social Context on Advertising Reception," *Journal of Marketing Research*, 44 (May), 284–96.

Rayner, Keith (1998), "Eye Movements in Reading and Information Processing: 20 Years of Research," *Psychological Bulletin*, 124 (3), 372–422.

———, C.M. Rotello, A.J. Stewart, J. Keir, and S.A. Duffy (2001), "Integrating Text and Pictorial Information: Eye Movements When Looking at Print Advertisements," *Journal of Experimental Psychology: Applied*, 7 (3), 219–26.

Reichle, Erik D., Keith Rayner, and Alexander Pollatsek (2003), "The E-Z Reader Model of Eye-Movement Control in Reading: Comparisons to Other Models," *Behavioral and Brain Sciences*, 26 (4), 445–526.

Roediger, Henry L. and Kathleen B. McDermott (2000), "Distortions of Memory," in *The Oxford Handbook of Memory*, Endel Tulving and Fergus I.M. Craik, eds. Oxford: Oxford University Press, 149–62.

Rossi, Peter E., Greg A. Allenby, and Rob McCulloch (2005), *Bayesian Statistics and Marketing*. New York: John Wiley & Sons.

Shepard, T. Mills (1942), "The Starch Application of the Recognition Technique," *Journal of Marketing*, 6 (April), 118–24.

Singh, Surendra N., Michael L. Rothschild, and Gilbert A. Churchill Jr. (1988), "Recognition Versus Recall as Measures of Television Commercial Forgetting," *Journal of Marketing Research*, 25 (February), 72–80.

Starch, Daniel (1923), *Principles of Advertising*. Chicago: A.W. Shaw Company.

Stuart, A. and J.K. Ord (1994), *Kendall's Advanced Theory of Statistics*, 6th ed. London: Edward Arnold.

Torralba, Antonio, Aude Oliva, Monica Castelhano, and John M. Henderson (2006), "Contextual Guidance of Eye Movements and Attention in Real-World Scenes: The Role of Global Features in Object Search," *Psychological Review*, 113 (4), 766–86.

Wedel, Michel and Rik Pieters (2000), "Eye Fixations on Advertisements and Memory for Brands: A Model and Findings," *Marketing Science*, 19 (4), 297–312.

——— and ——— (2007), "A Review of Eye-Tracking Research in Marketing," in *Review of Marketing Research*, Vol. 4, Naresh Malhotra, ed. Armonk, NY: M.E. Sharpe, 123–47.

Whittlesea, Bruce W.A. and Jason P. Leboe (2000), "The Heuristic Basis of Remembering and Classification: Fluency, Generation, and Resemblance," *Journal of Experimental Psychology: General*, 129 (1), 84–106.

Yonelinas, Andrew P. (2002), "The Nature of Recollection and Familiarity: A Review of 30 Years of Research," *Journal of Memory and Language*, 46 (3), 441–517.

Zhang, Jie, Michel Wedel, and Rik Pieters (2009), "Sales Effects of Attention to Feature Advertisements: A Bayesian Mediation Analysis," *Journal of Marketing Research*, 46 (October), 669–81.

Fonte: Journal of Marketing Research, v. 47, n. 3, p. 387-400, 2010. [Base de Dados]. Disponível em: <http://web.ebscohost.com>. Acesso em: 15 dez. 2010.