RAISING THE BAR:
BIAS-ADJUSTMENT OF ADVERTISING RECOGNITION TESTS

Anocha Aribarg
Stephen M. Ross School of Business
University of Michigan
701 Tappan Street
Ann Arbor, Michigan 48109-1234
Tele: (734) 763-0599
Fax: (734) 936-0279
E-mail: anocha@umich.edu

Rik Pieters
Department of Marketing
Faculty of Economics and Business Administration
University of Tilburg
P.O. Box 90153
5000 LE Tilburg
Tilburg, The Netherlands
Tele: +31 13 466-3256
Fax: +31 13 466-8354
Email: pieters@uvt.nl

Michel Wedel
Robert H. Smith School of Business
University of Maryland
3303 Van Munching Hall
College Park, Maryland 20742
Tele: (301) 405-2162
Fax: (301) 405-0146
E-mail: mwedel@rhsmith.umd.edu

RAISING THE BAR:
BIAS-ADJUSTMENT OF ADVERTISING RECOGNITION TESTS

Abstract

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

Ad recognition tests were pioneered by Daniel Starch (1923; Shepard 1942) and have been used ever since in marketing. In these tests, consumers report which ads they remember to have 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 read most of the copy in the ad (the “read-most” measure). Ad recognition tests provide measures of consumers’ direct memory for prior exposure to advertising. The more attention consumers have paid to the ad, brand and text, the higher the recognition scores in question are assumed to be. Although originally developed for print ads, recognition tests are also used to assess prior exposure to outdoor (Bhargava, Dontu, and Caron 1994), web (Havlena and Graham 2004) and television advertising (Heath and Nairn 2005; Mehta and Purvis 2006; Singh, Rothschild, and Churchill 1988), among others. Recognition scores have been popular metrics of ad effectiveness in advertising practice, where ad recognition is assessed after participants have been exposed to ads in their homes (Baldinger and Cook 2006; Belch and Belch 2001; Hanssens and Weitz 1980). They 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, using the ads as retrieval cues, and the resulting scores are readily comparable to benchmarks based on a long history of applications. These strengths contribute to their popularity.

Despite their extensive application, little is known about the accuracy of recognition tests as measures of attention to ads during prior exposure. This is surprising because memory research suggests that ad recognition may be systematically biased due to 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 measures 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 ads (Wedel and Pieters 2000), often measured by gaze durations (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 academia and practice, which makes it challenging to generalize the findings obtained under lab conditions to in-home situations. Given the prevalent use of recognition test in marketing academics to test advertising processing models, and in marketing practice to assess "which ads attract the most attention,"[1] and to guide advertising message and media decisions,[2] we believe it is imperative for our research to address these research questions.

Our research aims to make the following three contributions. First, we propose a new statistical model to examine the relationship between attention to print ads, as measured through eye-tracking methodology, and Starch ad recognition measures. The model accommodates the potential influence that the ad layout and the familiarity with the advertised brands have on attention and recognition memory. We observe that, as hypothesized, ad layout and brand familiarity indeed systematically biases 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 as the probabilities that consumers have actually seen a specific ad and its elements given that they claim recognition, and negative diagnostic values as the probabilities that consumers have actually *not* seen a specific ad and its elements given that they do *not* claim recognition. Significant differences in positive and negative diagnostic values across different recognition measures are revealed. In particular, the ad-noted measure has a high positive diagnostic value while the brand-associated measure has a high negative diagnostic value.

Third, we demonstrate how the positive and negative diagnostic values 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 ads are not available. Hold-out 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. The next section describes the data on which the analyses are based.

DATA

Data collection was done in cooperation with the market research agency Verify International (Netherlands). Four hundred and twenty eight consumers (50% females, age between 18 and 60) participated in the study for monetary compensation. Two hundred and forty three randomly-selected consumers from the participant pool of the market research agency received a copy of the latest issue of Cosmopolitan magazine containing all 48 full-page ads *at home* and were asked to use the magazine as they normally would, and come to the lab of the market research agency one week later, where they engaged in the ad recognition test. This situation mimics ad recognition testing in practice, although in practice the time delay between exposure and testing varies (the PARM study investigated the effect of the time delay and found modest effects of the delay for recognition; Bagozzi and Silk 1983).

The remaining 185 consumers were directly invited to the lab, where we collected data from them in three phases (described below). 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 on 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 for having abnormal vision.

Brand Familiarity. In phase 1, participants provided general information about their socio-demographics, and familiarity regarding a large set of products and brands (total *n* = 91), as well as about a number of 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 with “completely unknown” (score = 0), “unknown” (1), “known” (2), and “known very well” (3).

Eye-Tracking. In phase 2, attention to advertising was assessed 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 studied full-page advertisements), in fixed front-to-back order, while their eye-movements were recorded. They could inspect pages more closely if desired, 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 1,280 x 1,024 pixel resolution. Participants touched the lower-right corner of the (touch-sensitive) screen to proceed, as when leafing through print material.

Infrared corneal reflection methodology was used for eye tracking (Duchowski 2003). During data collection, participants could freely move their heads in a virtual box of about 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 of time during which the eye is relatively still and information uptake occurs. The duration of an individual fixation is around 200-400 ms (Rayner 1998). Gaze duration is the sum of individual fixation durations on an ad or its elements; both fixation frequencies and gaze durations on the ad 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 48 ads studied.

Ad Recognition. In phase 3, participants were exposed to each of the target ads from Cosmopolitan on a computer screen (after verifying that they remembered having seen this issue of the magazine; all had), and asked to indicate for each ad: “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), and in case of “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 similar to other ad recognition measures in ad theory and practice (Heath and Nairn 2005; Krishnan and Chakravarti 1999; Mehta and Purvis 2006). All ads were shown with their editorial counter-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 ads is shown during the test. Upon completion, participants were debriefed (none indicated to have expected the memory task when participating in the earlier phases of the study), thanked and paid. Table 1 gives summary statistics.

As expected, ad recognition scores, as percentage of participants answering “yes” to each of the measures, differed between the lab and in-home conditions. On average, 39.2% in the in-home condition indicated to recognize the ads, as compared to 54.3 % in the lab condition ($p <$ 0.05). Also, the brand-associated score was 29.5% in-home as compared to 40.5% in the lab ($p <$ 0.10). Unlike the ad-noted and brand-associated scores, scores for the read-most measure were close for the in-home (16.9%) and the lab (16.3%) conditions ($p >$ 0.10).

*** Insert Table 1 ***

Ad Content Analysis. Additional information about the ads, products and brands was collected through content analysis. A panel of 20 trained coders (10 male and 10 female graduate students) judged the ads and brands in individualized random order on eight seven-point rating scales. Scores were averaged per ad across judges (average alpha for the twenty coders was .892 across the eight items). A principal components analysis on the 8 ratings across the 48 target ads produced three clean components (with eigenvalues > 1), brand popularity, ad uniqueness, and ad attractiveness. Mean orthogonal component scores across items in the three scales are used in the post-hoc analyses. Brand popularity comprised of three items: (a) “I know this brand …,” from 1 not at all to 7 very well, (b) “I have seen this specific advertisement for this brand…”

from 1 never before to 7 very often, and (c) “I am … with this brand,” from 1 not at all familiar to 7 very familiar. Ad uniqueness comprised of three items: “To me, this specific advertisement for this brand is …,” (a) 1 not at all unique to 7 very unique, (b) 1 not at all original to 7 very original, and (c) 1 not at all unexpected to 7 very unexpected. Finally, ad attractiveness comprised of two items: (a) 1 not at all attractive to 7 very attractive, and (b) 1 not at all exciting to 7 very exciting. In addition, the number of words in the headline was counted (mean = 4.67, std = 2.71) because of its potential influence on attention to advertising (Rayner, Rotello, Stewart, Keir, and Duffy 2001).

A MODEL OF ATTENTION AND AD RECOGNITION

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

Attention Model

We have $l=1,..., L$ ads, each consisting of $j=1, ..., J$ ad design elements, a sample of $i=1, ..., I$ consumers, and $m=1, ..., M$ recognition measures. There are $J = 3$ ad design elements, that is, pictorial, text, and brand, and $M = 3$ recognition measures, that is, 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, $t_{i,j,l}$, 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 $i = 1,..,n$ 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), while 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). We thus model gaze duration on a specific element as the sum of the durations of the individual fixations on that element: $t_{i,j,l} = \sum_{k_i=1}^{n_{i,j,l}} d_{k_{i,j,l}}$, with $d_{k_i,j,l}$ is duration of the kth fixation of the i-th 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 the marginal distribution of fixation frequency ($n_{i,j,l}$) to be Poisson (Wedel and Pieters 2000) with parameter $\mu_{i,j,l}$ and the distribution of fixation durations ($d_{k_i,j,l}$) to be exponential with parameter $\lambda_{i,j,l}$ (Harris et al. 1988). Conditional upon $n_{i,j,l}$, $t_{i,j,l}$ 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:

$$(1) \qquad \textit{Fixations: } f_N(n_{i,j,l}) = Poisson(n_{i,j,l} \mid \mu_{i,j,l}) = \frac{\mu_{i,j,l}^{n_{i,j,l}} \exp[-\mu_{i,j,l}]}{n_{i,j,l}!}$$

$$\textit{Gaze: } f_{T|N}(t_{i,j,l} \mid n_{i,j,l}) = Gamma\left(t_{i,j,l} \mid n_{i,j,l}, \lambda_{i,j,l}\right) = t_{i,j,l}^{n_{i,j,l}-1} \frac{\exp[-t_{i,j,l} / \lambda_{i,j,l}]}{\lambda_{i,j,l}^{n_{i,j,l}} \Gamma(n_{i,j,l})}$$

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):

$$(2) \qquad \mu_{i,j,l} = \exp\left(x_{i,j,l}^{A}{}' \alpha_{i,j} \right), \text{ and } \alpha_{i,1:J} \sim MVN\left(\overline{\alpha}, D_{\alpha}\right)$$

$$\lambda_{i,j,l} = \exp(\lambda_{i,j,l}^{*}) \text{, and } \lambda_{i,1:J}^{*} \sim MVN(\overline{\lambda}, D_{\lambda})$$

where $\alpha_{i,1:J} \equiv vec(\alpha_i)$, with α_i a $(J \times P^A)$ matrix with P^A 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 shown in (2), to account for heterogeneity among consumers and over-dispersion of the fixation counts. Thus, this model-component extends the multivariate Poisson log-normal distribution (Chib and Winkelman 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 (Wedel and Pieters 2000; Pieters and Wedel 2004). Brand familiarity (as a person-related factor) is also predicted to influence attention to the ad and its elements (Reichle, Rayner, and Pollatsek 2003). Therefore, all these variables are included 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

We have the binary variables indicating 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 or not s/he remembers to have seen the ad, and if so, whether or not s/he remembers to have identified the brand, and to have read most of the text. Attention is assumed to be unobserved, but reflected in the total gaze duration (Rayner 1998). Recognition is claimed 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:

(3) Ad-noted: $f_Y(y_{i,1,l} \mid \omega_{i,1,l}) = P(y_{i,1,l} = 1 \mid \omega_{i,1,l})^{y_{i,1,l}} \, P(y_{i,1,l} = 0 \mid \omega_{i,1,l})^{1-y_{i,1,l}}$

Brand-associated and read-most: $f_Y(y_{i,m,l} \mid y_{i,1,l} = 1, \omega_{i,m,l}) =$

$$P(y_{i,1,l} = 1 \mid \omega_{i,1,l}) P(y_{i,m,l} = 1 \mid y_{i,1,l} = 1, \omega_{i,m,l})^{y_{i,m,l}} \, P(y_{i,m,l} = 0 \mid y_{i,m,l} = 1, \omega_{i,m,l})^{1-y_{i,m,l}}, m = 2,3$$

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

$$\text{(4)} \qquad \omega_{i,1,l} = \beta_{i,1,0} + \phi_{i,l}\,\beta_{i,1,\phi} + x_{i,1,l}^{M\prime}\,\beta_{i,1}, \text{ and } \beta_i \equiv (\beta_{i,1,0}, \beta'_{i,1,\phi}, \beta'_{i,1})' \sim MVN(\bar{\beta}, D_\beta)$$

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

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

where the explanatory variables are $\phi_{i,l}$, attention to each of the three ad-elements as explained in detail below, and $x_{i,m,l}^{M}$, the size of the ad elements and brand familiarity. The parameters β_i and γ_i follow multivariate normal distributions, as shown in (4), to account for heterogeneity among consumers. Note that we assume the individual-level parameters to be uncorrelated across equations (2) and (4).[3,4]

The probability that a consumer claims to have noted the ad (m = 1), identified the brand (m = 2) and read most of the text (m = 3) are modeled as a function of attention to the ad 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). Henderson (1992), for example, describes their relation through a "rubber-band" metaphor, with the eyes and attention closely but imperfectly coupled. We therefore assume gaze duration on an ad element for a specific ad to be 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 $\phi_{i,l} = \left(\mu_{i,1,l}\lambda_{i,1,l}, \mu_{i,2,l}\lambda_{i,2,l}, \mu_{i,3,l}\lambda_{i,3,l}\right)$ 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, $\gamma'_{i,m,\phi}$ captures the effects of attention on brand-associated ($m=2$) and read-most ($m=3$). Recognition is claimed 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 Wedel and Pieters (2000), who include fixation frequencies, rather than unobserved attention, in a binary probit memory model.

Because the original ad is available to the participants during the recognition test, we predict the sizes of the three ad elements to 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 ad and its elements. For example, a large pictorial element may lead consumers to infer that they must have seen the ad, 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 ad 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 ads for familiar brands. We will not be 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, over and above their indirect effects mediated through attention to the ad or ad element.

Thus, the model specified in equations (1) through (4) allows for tests of the effects of ad layout on attention, their indirect effects on ad recognition mediated by attention (MacKinnon, Fairchild, and Fritz 2007), and their direct effects on recognition over and above their effects via attention. These latter direct effects would demonstrate systematic biases in the recognition scores, which 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, a Metropolis-within-Gibbs sampling algorithm is used to estimate the model (Rossi, Allenby, and McCulloch 2005). Multivariate normal priors are used for all regression coefficients with mean zero and variance $10^4 I$. 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 proposed full model with a set of simpler alternatives to gain insight into the contribution of each of the specific model components, based on their log-marginal densities (LMD). To compute the log-marginal densities, 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 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 ads and other factors. Therefore, it can be used deductively to establish the probability that recognition is claimed when consumers attended to the ad and its elements, and the probability that recognition is not claimed when consumers did not attend to the ad 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 ads during prior exposure in situations where attention is not directly measured, through for example eye-tracking. In those applications one 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 has been defined as the proportion of people with positive test results who are accurately diagnosed to have the condition in question, and the negative predictive value as the proportion of people with negative test results who are accurately diagnosed to not have it (Altman and Bland 1994b; Phelps and Ghaemi 2006).

We propose to assess the diagnosticity of recognition tests through the *positive diagnostic value* (PDV) and the *negative diagnostic value* (NDV), and develop a procedure that provides bias-adjusted recognition measures based on these metrics. We define PDV as the probability that during exposure an individual has fixated on an ad or a specific ad element *at least* a certain number of times or more, given that s/he claims to have seen it. Similarly, we define NDV as the probability that an individual has fixated on an ad or a specific ad element *at most* a certain number of times, given that s/he claims to *not* have seen it. These diagnostic values are thus the *inverse* conditional probabilities of fixating on an ad or element conditional upon claimed recognition of the ad or element. Bayes theorem can be used to derive these predictive values (Goodman 1999).

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, as the PDV of the recognition test. We similarly 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:

$$\begin{aligned} &\mathrm{PDV}(\chi) = p\left(n_{i,j,\ell} \geq \chi_{PDV} \mid I_{m,l} = 1\right) \\ &\mathrm{NDV}(\chi) = p\left(n_{i,j,\ell} < \chi_{NDV} \mid I_{m,l} = 0\right). \end{aligned} \tag{5}$$

Equation (5) can be evaluated based on the parameter estimates obtained from the attention and memory model using Bayes' theorem. That is, the inverse probability that an individual has fixated on the ad or ad element, in case s/he claims (no) recognition, $p(N_{i,j,l}=n_{i,j,l}| y_{i,m,l})$ is computed as:

$$\frac{\int\limits_{\omega}\int\limits_{\lambda}\int\limits_{\mu}\int\limits_{t} f_N\left(n_{i,j,l} \mid \mu_{i,j,l}\right) f_{T|N}(t_{i,t,j} \mid n_{i,j,l}, \lambda_{i,j,l}) f_Y\left(y_{i,m,l} \mid \omega_{i,m,l}\right) p(\mu_{i,j,l}, \lambda_{i,j,l}, \omega_{i,m,l} \mid N, T, Y) \partial t \partial \mu \partial \lambda \partial \omega}{\int\limits_{\omega}\int\limits_{\lambda}\int\limits_{\mu} p\left(y_{i,m,l} \mid \omega_{i,m,l}\right) p(\mu_{i,j,l}, \lambda_{i,j,l}, \omega_{i,m,l} \mid N, T, Y) d\mu d\lambda d\omega} \tag{6}$$

Here $f_N\left(n_{i,j,\ell} \mid \mu_{i,j,l}\right)$ is the conditional probability of observing $n_{i,j,l}$ fixations given $\mu_{i,j,l}$, and $f_{T|N}\left(t_{i,j,l} \mid n_{i,j,l}, \lambda_{i,j,l}\right)$ 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*. $f_Y\left(y_{i,m,l} = 1 \mid \omega_{i,m,l}\right)$ 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; $m = 3$ for read-most), given his/her latent attention $\phi_{i,\ell}$ 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, $\phi_{i,\ell}$. 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, read-most) for each ad, in the MCMC chain after the burn-in period, we first compute $f_N(n_{i,j,l} < \chi \mid \mu_{i,j,l})$ and $f_Y(y_{i,m,l} = 1 \mid \omega_{i,m,l})$ based on Equations (1) and (3), respectively. The term $f_Y(y_{i,m,l} = 1 \mid \omega_{i,m,l})$ is used for the denominator of the PDV and $1 - f_Y(y_{i,m,l} = 1 \mid \omega_{i,m,l})$ for the denominator of the NDV. Next, we compute $(1 - f_N(n_{i,j,l} < \chi \mid \mu_{i,j,l})) \times f_Y(y_{i,m,l} = 1 \mid \omega_{i,m,l})$ for the numerator of the PDV and $(f_N(n_{i,j,l} < \chi \mid \mu_{i,j,l}) \times (1 - f_Y(y_{i,m,l} = 1 \mid \omega_{i,m,l}))$ for the numerator of the NDV. After the MCMC run, we average numerator draws and then denominator draws for each ad 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 is for a specific threshold χ_{PDV}, the more diagnostic the recognition measure is for prior attention to the ad 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 ad 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 Sullivan 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 will demonstrate how these diagnostic values can be used in a bias-adjustment procedure for ad recognition measures.

RESULTS

We compare the log-marginal density (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 start with a baseline model containing 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, over and above its effects mediated by attention. 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, over and above 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 due to memory retrieval factors. We present parameter estimates of the third model.[5]

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 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, Pollatsek, and Rayner 2003), whereas the gist of pictorials can often be grasped in a glance (Rayner 1998). Ad elements generally compete for attention, as shown by significant negative cross-effects of their sizes, for instance larger pictorial sizes reducing 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 ads.

*** Insert Table 2 ***

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 ads. Table 3 also shows that ad layout has direct effects on recognition memory, over and above those mediated by attention. A larger pictorial increases ad noted, regardless of how much attention was devoted to the ad during the earlier exposure. Our finding is consistent with findings on the effect of pictorial size on ad recognition measures (Finn 1988), but ours shows the effect to be independent of the actual attention devoted to the pictorial. Thus, larger pictorials in ads lead to a systematic over claiming of prior attention to the ads.

In addition, more text in the ad 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, although 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.[6] 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 ads for familiar brands more often and to have read most of their text, independent of their actual attention to the ads. This along with the finding that familiar ads receive more fixations on the pictorial and text, but not the brand, may indicate that familiarity with the brand lowers the threshold for ad recognition.

Taken together, these results reveal that whereas recognition memory for the ad as a whole and its brand element reflect prior attention to some extent, memory for text is mostly reconstructed during the recognition test and bears little relation with 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.

*** Insert Table 3 ***

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 ads and consumers, against values of the fixation threshold (χ =0,1,2, …). Figure 1 also depicts the interval containing 90% of the ads, for each of these curves.[7] In interpreting the PDV and NDV and de-biasing the recognition scores, we focus on $\chi_{PDV} = \chi_{NDV}$ =5. Although other thresholds are readily accommodated, five fixations are a natural cut-off in eye-tracking studies of complex scenes such as ads, and have been used in a range of studies by, amongst others, Charness, Reingold, Pomplun, and Stampe (2001), Masciochi, Mihalas, Parkhurst, and Niebur (2008), Torralba, Oliva, Castelhano, and Henderson (2006). This threshold corresponds to roughly 1-2 seconds of exposure needed for reliable recognition memory, which reflects exposure durations to ads in natural conditions for the majority of people (Pieters and Wedel 2004). We have 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. We therefore believe that $\chi_{PDV} = \chi_{NDV}$ =5 will be a reasonable choice in many studies. Its validity is further investigated below.

*** Insert Figure 1 ***

The top panel of Figure 1 shows that the ad-noted measure has the highest positive diagnostic value. If consumers claim to recognize the ad (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, respectively 26.2 and 35.8%. To illustrate, at the threshold, the odds of the PDVs of ad-noted over brand-associated are almost 4:1 (0.93/0.26) in favor of ad-noted. Note, however, that the number of fixations on the brand and text are smaller than those on the ad as a whole (Table 1).

The bottom panel of Figure 1 shows that the brand-associated measure has the highest negative diagnostic value. If consumers claim not to have noted the brand in the ad (NDV brand-associated), the probability of having had less than five fixations during the original ad exposure is 81.6%. If they claim not to have read-most (NDV read-most), the probability of having had less than five fixations is 71.8%, which is also fairly high. However, if consumers claim not to have noted the ad (NDV ad-noted), the probability of having had less than five fixations is only 12.2%. This suggests that claims to not have noted the ads are unreliable and that false negative claims are very common as long as consumers fixate on an ad less than 1-2 seconds. To illustrate, at the fixation threshold, the odds of brand-associated over ad-noted NDV is close to 7:1 (0.82/0.12) in favor of brand-associated.

Thus, the ad-noted measure has the highest positive diagnostic value, but at the same time the lowest negative diagnostic value, while the reverse holds for the brand-associated measure. If consumers claim to have noted an ad, there is high probability (92.6%) that they fixated the ad at least 5 times, and if they claim to not have noted the brand in the ad, there is fairly high probability (81.6%) that they fixated it less than 5 times.

Based on this, 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 ad and its elements, even when they may not actually have attended them. In situations where these raw recognition scores are available, but eye-tracking data are not, it may be useful to be able to adjust the raw scores to remove biases. The bias-adjusted recognition (BAR) scores indicate the estimated proportion of consumers who have fixated on the ad or its elements five times or more. Our proposed adjustment uses values of $\mathrm{PDV}(\chi)$ and $\mathrm{NDV}(\chi)$ that can be read directly from Figure 1 at the threshold $\chi=5$ (or any other desired threshold). Bias-adjusted recognition scores can be computed as follows:

(7) $$BAR\ Score = \mathrm{PDV}(\chi) \times \{\text{Raw Score}\} + (1 - \mathrm{NDV}(\chi)) \times \{1 - \text{Raw Score}\}.$$

Equation (7) is derived from the rule of total probability: $P(A)=P(A|B)P(B)+P(A \mid \overline{B})P(\overline{B})$. Here, $P(A)$ is the quantity required but unknown from a recognition test: the probability that consumers fixate on an ad (or the brand or text element) five times or more (the BAR score). $P(A|B)$ is the probability that consumers fixate on the ad five times or more, given claimed recognition, which is the PDV. $P(B)$ is the probability that consumers claim ad recognition (raw recognition score). $P(A \mid \overline{B})$ is the probability that consumers fixate on the ad five times or more, given no claimed recognition, which equals (1-NDV). Finally, $P(\overline{B})$ 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 ads and consumers for which only the recognition--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, then 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. Holding all other things 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.

In order to investigate the diagnostic values further, we regressed the log-odds diagnosticity (log[PDV/(1-NDV)]) for each ad on its associated brand popularity, ad-attractiveness, and ad-uniqueness ratings, and 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 ads 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, but ad attractiveness a positive effect on the diagnosticity of the read-most measure.

Hold-out Validation. To demonstrate the improved accuracy of BAR scores over raw recognition scores, we use two hold-out samples. First we re-estimate the model for all participants and a random sample of 38 ads, and retain 10 ads as a hold-out sample. Second, we re-estimate the model for a random sample of 38 ads and 145 participants, and retain 10 ads and 40 participants as a hold-out sample. The first hold-out sample enables us to assess the performance of our approach for a sample of new ads for the same participants in the test; the second hold-out sample allows us to assess performance for a new sample of ads and a new sample of participants. We adjust the raw scores of the hold-out sample of ads using equation (7), with PDV and NDV estimated from the calibration sample, averaging PDV and NDV across participants and ads for each of the recognition measures. We define the true score as the proportion of consumers who actually fixated on the ad 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 ad. Averaging these absolute deviations across the ads in the hold-out sample, we obtain the mean absolute deviations of the raw (MAD_r) and bias-adjusted recognition (MAD_b) scores. Table 4 gives the in-sample and out-of-sample results.

*** Insert Table 4 ***

As expected, BAR scores are more accurate than the raw scores in reflecting actual fixations on the ad 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 ads/same participants, and of new participants/new ads 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 around 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 for the ad-noted score is very close to the in-sample MAD, 9.8%, the out-of-sample MADs are even somewhat smaller for the brand-associated (10.1%) and read-most scores (14.0%). This may have been due to the specific ads in our (random) hold-out sample. The magnitudes of these out-of-sample MADs (new ads, new participants) are indicative of good performance of the bias adjustment procedure. We also compute the percentage improvement in bias-adjusted recognition scores relative to MAD_r (Table 4). Improvement in accuracy ranges from roughly 25% to 60% out-of-sample, which is substantial.[8]

Bias-adjustment for Ad Recognition In-home. So far the results were obtained from data collected in a laboratory setting, because only there could eye-movements and recognition measures be collected from the same people. Yet, bias-adjustment seems particularly valuable when ad recognition testing takes place under natural exposure conditions which is common, where attention to ads is short, in the order of magnitude of a few seconds (Pieters and Wedel 2004), and where eye-tracking measures are typically unavailable. We therefore 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 home condition. To explore the effects of bias-adjustment in this setting, we compare the recognition scores between the home and lab conditions before and after bias-adjustment. Participants were randomly allocated to one of the two conditions, and the same set of ads was evaluated in the same editorial context. If indeed common retrieval biases would be removed by the bias-adjustment procedure, then the scores between in-home and lab conditions should be closer after our correction.

In Table 5, the differences between the raw recognition scores in the in-home and 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 to true scores than are the raw scores. After bias-correction, however, the differences between the home and lab conditions diminished substantially, to .7%, .6% and .5% respectively. These results are in part due to the low diagnosticity of the test, but, they do reveal that bias-adjustment reduces the gap between the recognition scores of the in-home and lab conditions, and corrects recognition scores collected after exposure in natural in-home settings, as frequently used in practice. This supports the potential improvements due to the proposed bias-adjustment procedure.

DISCUSSION

Diagnosticity of Ad-recognition Scores

We found that attention to the ad predicted the ad-noted measure, and 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 ads did not significantly affect the read-most measure. Independent of attention, consumers over-claimed ad recognition when the ad 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, is a typical ad layout. Thus regardless of actual attention to them during prior exposure, recognition of ads with prototypical layouts was over-claimed in recognition tests. Failure to control for ad-prototypicality in ad recognition measures may lead to overrating the effectiveness of the specific ads.

The diagnostic value of ad recognition is low, but varies across measures and metrics. Specifically, the positive diagnostic value of the ad-noted measure was high, but its negative diagnostic value was low, so that the ad-noted measure was best at identifying ads that were actually noted. Conversely, the brand associated and read-most scores had lower positive diagnostic values, but higher negative diagnostic values, so that the brand-associated and read-most recognition measures were better at excluding ads for which the brand element was actually not identified and the text not read. None of the ad recognition measures performed well in both accurately identifying attended and excluding unattended ads. An interesting question for future 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, supporting evidence, using "consider the opposite" strategies, or cueing de-biasing factors (Arkes 1991). Triangulation with other memory measures, such as recall and indirect measures of memory is another viable route (Krishnan and Chakravarti 1999).

We proposed bias-adjusted recognition (BAR) scores that indicate the proportion of consumers who have fixated on the ad 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 those 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 ads.

Because we could not track consumers' eye-movements at home, we were not able to assess biases for that condition directly. But, because the memory traces in the 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 our results on the reduction of biases in the home-condition may be consistent with the presence of common retrieval biases, other, more mechanical, explanations cannot be excluded. Future research may address these issues.

The threshold of five fixations, which was used for each of the three ad-recognition measures, was chosen based on 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 positive and negative diagnostic value, with two as theoretical maximum) never exceeded 1.14 for any threshold value, which is low. For the brand-associated measure, there was no threshold where the positive and negative diagnostic values 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 positive and negative diagnostic values ad recognition tests are still useful. This, as well as the optimal choice of the recognition thresholds is an important topic for future research.

Implications for Theory and Practice

In academic advertising research, the use of ad recognition measures may misdirect theory development. In the present study, for instance, larger pictorials increased the ad-noted measure substantially, independent of the actual attention devoted to the ad. This may lead to overvaluing 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 instance, in our 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 post-testing and campaign evaluation, with some practitioners even calling them “the definitive advertising measurement scores.”[9] Thus, ad-noted, brand-associated and read-most scores across magazines and product categories have been used to benchmark the effectiveness of print advertising[10], and similar recognition measures are used in television advertising.[11] They are being used to assess which ads attract most attention, and 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: they are not strong proxies for attention and in particular text recognition is not related to attention at all. Memory biases may especially harm prototypical ads, because their ad-noted and read-most scores tend to be over-valued, independent of actual attention during exposure. Comforted by high recognition scores, ads may then be insufficiently optimized and their campaigns may be sustained beyond the cost-effective level of repeated exposures. Benchmarking ads against other ads based on raw ad recognition measures requires caution, given the wide variations in positive and negative diagnostic values across ads (See the Web-Appendix). One reason why 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 present findings indicate that they may unfortunately have insufficient diagnostic value for this purpose.

Although this research focused on diagnosticity of Starch recognition tests for print ads, 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 past exposure, can attempts at improving their future performance become effective. The proposed framework for diagnosticity and bias adjustment of recognition tests hopes to contribute to such 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.

——— 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, 487–505.

Bhargava, Mukesh, Naveen Donthu, and Rosanne Caron (1994), “Improving the Effectiveness of Outdoor Advertising,” *Journal of Advertising Research*, 34 (2), 46–55.

Belch, George E. and Michael A. Belch (2001), *Advertising and Promotion: An Integrated Marketing Communications Perspective*, 5th ed. Boston: McGraw-Hill.

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.

Ehrenberg, Andrew S.C., Gerald J. Goodhardt, and T. Patrick Barwise (1990), "Double Jeopardy Revisited," *Journal of Marketing*, 54 (3), 82–91.

Finn, Adam (1988), "Print Ad Recognition Readership Scores: An Information Processing Perspective," *Journal of Marketing Research*, 25 (2), 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 (3), 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.

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*, Vol. I and II. 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, 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, 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.

Rayner, Keith, 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 and Sons.

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

Singh, Surendra N. and Gilbert A. Churchill Jr. (1986), "Using the Theory of Signal Detection to

Improve Ad Recognition Testing," *Journal of Marketing Research*, 23 (4), 327–36.

———, Michael L. Rotschild, and Gilbert A. Churchill Jr. (1988), "Recognition Versus Recall as Measures of Television Commercial Forgetting," *Journal of Marketing Research*, 25 (1), 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, Wagner Kamakura, Neeraj Arora, Albert Bemmaor, Jeongwen Chiang, Terry Elrod, Rich Johnson, Peter Lenk, Scott Neslin, and Carsten Stig Poulsen (1999), "Discrete and Continuous Representations of Unobserved Heterogeneity in Choice Modeling," *Marketing Letters*, 10 (3) 219–32.

——— 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," *Review of Marketing Research*, 4, 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 Feature Advertisements: A Bayesian Mediation Analysis," *Journal of Marketing Research*, forthcoming, 46 (October).

1. http://www.time.com/time/mediakit/audience/research/proprietary/starch.html, accessed March 2009.
2. See Hermie, Patrick, Trui Lankcriet, Koen Lansloot, and Stef Peeters (2005), *StopWatch. Everything of the Impact of Advertisements in Magazines*, Diegem, Belgium: Sanoma Magazines, accessed from http://www.ppamarketing.net/cgi-bin/go.pl/research/article.html?uid=116, March 2009.
3. We have allowed the error terms of the model components m=1,2,3 to co-vary, based on a suggestion by Peter Lenk. These covariances were not significant in the application and the model yielded very similar estimates.
4. We have parameterized the heterogeneity distribution as a function of gender, but did not find significant effects in the application and will not report these effects.
5. We also estimated a version of the memory model in which we include observed fixations, rather than latent attention, and find that the estimates of the two models are comparable. The full model is preferable on theoretical grounds.
6. 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.
7. The Web-appendix provides the diagnostic value curves separately for each of the 48 ads in our study.
8. We also corrected Starch scores using PDV and NDV computed simply as the proportion of participants who had fixated on an ad greater than or equal to (less than) the fixation threshold given that they reported to have (not) seen the ad. This correction leads to worse prediction than raw Starch scores.
9. http://www.mcnairingenuity.com.au, accessed July 2008.
10. http://findarticles.com/p/articles/mi_m4PRN/is_2008_June_3/ai_n25475031, http://www.gfkamerica.com/practice_areas/brand_and_comm/starch/adnorms/index.en.html, accessed July 2008.
11. www.ameritest.net/products/adtracking.pdf, accessed July 2008.

TABLE 1

DESCRIPTIVE STATISTICS

Variable	N	Mean	SD	Median	Min	Max
Ads: Surface sizes:						
Brand (*inch*^{2})	48	11.134	8.442	8.168	1.825	45.752
Pictorial (inch^2)	48	64.947	16.229	68.536	8.791	81.632
Text (*inch*^{2})	48	15.499	11.575	16.792	0.000	46.938
Laboratory group (*n* = 185):						
Brand familiarity (0,...,4)	8880	1.873	0.983	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 (*sec.*)	8880	0.605	0.765	0.38	0	9.12
Pictorial (*sec.*)	8880	1.173	1.117	0.86	0	14.48
Text (*sec.*)	8880	0.811	1.303	0.34	0	17.02
Total (*sec.*)	8880	2.589	2.469	1.92	0	26.22
Recognition memory:						
Ad noted (0,...,1)	8880	0.543	0.498			
Brand associated (0,...,1)	8880	0.405	0.491			
Read most (0,...,1)	8880	0.163	0.369			
In-home group (*n* = 243):						
Recognition memory:						
Ad noted (0,...,1)	11664	0.392	0.488			
Brand associated (0,...,1)	11664	0.295	0.456			
Read most (0,...,1)	11664	0.169	0.375			

Note - Mean values of recognition memory measures are proportions.

TABLE 2
DETERMINANTS OF AD ATTENTION

	Attention to advertising					
	Brand		Pictorial		Text	
Predictors	Mean	SD	Mean	SD	Mean	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

Note - Bolded parameter estimates indicate that probabilities of the parameters to be larger or smaller than zero are greater than .95. Correlations are between parentheses.

TABLE 3
DETERMINANTS OF AD RECOGNITION

	Ad recognition memory					
	Ad noted		Brand associated		Read most	
Predictors	Mean	SD	Mean	SD	Mean	SD
Intercept	**-.636**	.112	**-.569**	.095	**-.633**	.101
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

Note - Bolded 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)

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

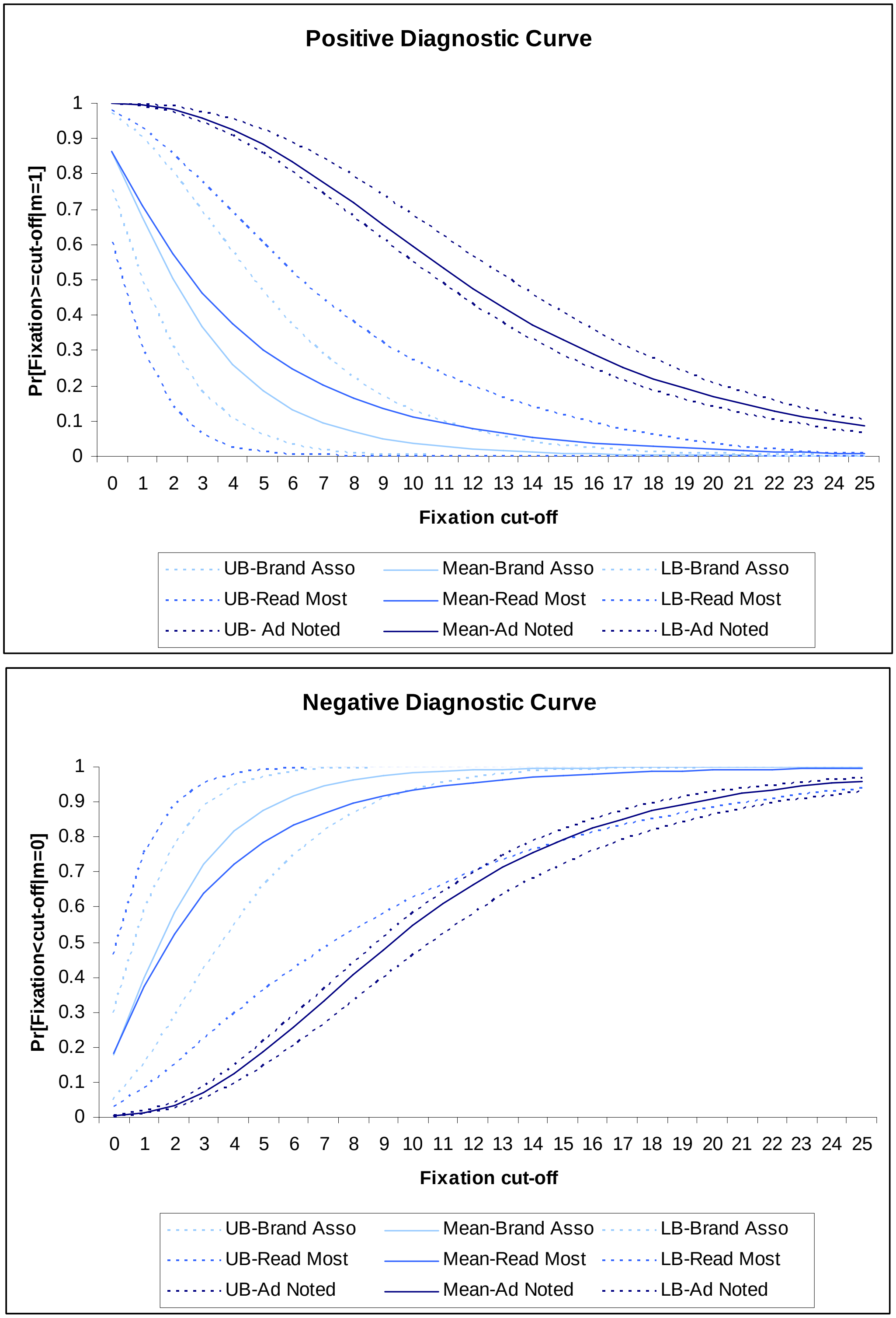
Note - % improvement in MAD = $(\text{MAD}_r - \text{MAD}_b)/\text{MAD}_r \times 100$

- Bolded (bolded and italicized) parameter estimates indicate statistical significance at α = .05 (.10).

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

FIGURE 1
POSITIVE AND NEGATIVE DIAGNOSTICITY CURVES
(positive (pr[number fixation>=fixation cut-off | m=1]) and negative (pr[number fixation<fixation cut-off | m=0]) diagnostic values at different fixation thresholds)



Positive Diagnostic Curve

Negative Diagnostic Curve