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. First, using a new statistical model, we establish the predictive value of consumers' actual attention during print ad exposure, as assessed through eye tracking, for subsequent ad recognition. We find ad recognition 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 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 measures 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 noticed 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, using the ads as visual retrieval cues. Although originally developed for print ads, recognition tests are also used to assess prior exposure to outdoor, web and television advertising, among others (e.g., Havlena and Graham 2004; Heath and Nairn 2005; Singh, Rothschild, and Churchill 1988). They have been popular as metrics of ad effectiveness in advertising practice, in which 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, in which participants are mostly exposed to ads under more controlled laboratory conditions (Bagozzi and Silk 1980; Finn 1988, 1992; Mothersbaugh, Huhmann, and Franke 2002; Puntoni and Tavassoli 2007). Ad recognition tests are easy to administer and the resulting ad performance scores are readily comparable to benchmarks based on a long history of applications, which are strengths partially explaining their popularity.

Despite their wide 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 tests may be systematically biased due to memory reconstruction processes during retrieval (Johnson, Hashtroudi, and Lindsay 1993; Mitchell and Johnson 2000; Roediger and McDermott 2000; Yonelinas 2002). That 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 developing advertising theory. Yet, direct tests of the diagnostic value of recognition tests are as yet unavailable. Also, little is known about the stability of recognition measures across the different exposure conditions used in academia and practice. This makes it challenging to generalize the findings obtained under lab conditions to in-home situations.

The present study aims to make the following three contributions. First, it establishes the *predictive* value of consumers' attention to advertising at exposure for the scores that the ads attain in subsequent recognition tests. To this aim, we propose a new statistical model of the relation between attention to print ads, as measured through eye-tracking methodology, and Starch ad recognition measures. The model accommodates the potential influence that the layout of ads and the familiarity with the advertised brands have on respectively attention and recognition memory. We observe that, as hypothesized, ad layout and brand familiarity indeed systematically bias ad recognition measures, independent of their effect 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, i.e., the probabilities that consumers have actually seen a specific ad and its elements given that they claim recognition, and negative diagnostic values, i.e., the probabilities that consumers have actually not seen a specific ad and its elements given that they do not claim recognition. This reveals significant differences in positive and negative diagnostic values between recognition measures, in particular for ad-noted and brand-associated.

Third, we demonstrate how the positive and negative diagnostic values of ad recognition measures can be used to develop bias-adjusted recognition (BAR) scores. Hold-out validation tests show that our bias-adjustment procedure substantially improves the diagnostic value of ad recognition measures. We assess the stability of recognition measures across in-home and laboratory testing conditions and apply the bias-adjustment procedure to recognition in both conditions. The results reveal that our procedure helps mitigate the differences in the measures obtained from these two conditions. The next section first 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 forty three randomly-selected consumers from the participant pool of the market research agency received a copy of the latest (December) issue of Cosmopolitan magazine containing 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 several unrelated tasks, as well as the ad recognition test. This situation mimics ad recognition testing in practice. The remaining 185 consumers were directly invited to the lab. Data collection for those participants was in three phases, with unrelated filler tasks in-between to clear memory. In phase 1, general information about the participants was collected; in phase 2, eye tracking was conducted to obtain measures of attention to advertising, and in phase 3, the ad recognition test was administered. Participants in the in-home condition engaged in the same ad recognition test as

those in the lab condition (phase 3). Participants were not a-priori made aware of the ad recognition test in phase 3.

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 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 48 full-page advertisements) while their eye-movements were recorded. They could inspect pages more closely if desired, as when exploring a magazine at home (Janiszewski 1998). 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. 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 brand, text, and pictorial as the main ad design elements were retained for each of the 185 participants and 48 target ads.

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). All ads were shown with their editorial counter-page, and in the order that 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. Also,

the brand-associated score was 29.5% in-home as compared to 40.5% in the lab. 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.

*** Insert Table 1 ***

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 185 participants in the lab condition.

We have $\ell = 1,...,L$ 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, $t_{i,j,\ell}$, the fixation frequency of consumer *i* on element *j* of ad ℓ, $n_{i,j,l}$ and the binary variables indicating "yes" or "no" to recognition measure *m* for consumer *i* for ad ℓ, $y_{i,m,\ell}$. The proposed attention component describes gaze duration as the sum of individual fixation durations through a hierarchical randomly stopped 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). Specifically, we model gaze duration on a specific element as the sum of the durations of the individual fixations on that element: $t_{i,j,l} = \sum_{k_i=1}^{n_{i,j,\ell}} d_{k_{i,j,l}}$, with $d_{k_i,j,\ell}$ is duration of the k^{th} fixation, and

further assume fixation duration $d_{k_i,j,l} = d_{i,j,l}; \forall k$. This defines the distribution of $t_{i,j,l}$ as a randomly stopped sum (Heller, Stasinopoulos, Rigby, and de Jong 2007; Johnson, Kotz, and Balakrishnan 1994; Stuart and Ord 1994). 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. We assume the marginal distribution of fixation frequency ($n_{i,j,l}$) to be Poisson (Wedel and Pieters 2000) and the fixation duration ($d_{i,j,\ell}$) to be Gamma distributed (Harris, Hainline, Abramov, Lemerise, and Camenzuli 1988).

The recognition memory component is a multivariate Probit model (Edward and Allenby 2003; Manchanda, Ansari and Gupta 1999), in which attention is specified to affect multiple correlated memory measures. 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.

Thus, attention and recognition memory for consumer *i*, ad element *j*, ad *l* and recognition measure *m* are:

(1) *Attention:* $f_N(n_{i,j,l}) = Poisson(n_{i,j,l} \mid \mu_{i,j,l})$

Attention: $f_{T|N}(t_{i,j,l} \mid n_{i,j,l}) = f(d_{i,j,l} \mid n_{i,j,l}\lambda_{i,j,l}, n_{i,j,l}^{-1}\rho_j) = Gamma(d_{i,j,l} \mid n_{i,j,l}, \lambda_{i,j,l}, \rho_j)$

Recognition memory: $f_Y(y_{i,m,l} \mid \mu_{i,j,l}, \lambda_{i,j,l}) = Bernoulli(y_{i,m,l} \mid \omega_{i,m,l})$

Expected fixation frequency $\mu_{i,j,\ell}$, and expected memory $\omega_{i,m,\ell}$, are 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) $$\mu_{i,j,\ell} = \exp\left(x^{A}_{i,j,\ell}{}' \alpha_{i,j} \right), \text{ and } \alpha_{i,1:J} \sim MVN(\overline{\alpha}, D_{\alpha}),$$

(3) $$\omega_{i,m,\ell} = \beta_{i,m,0} + \phi_{i,m,\ell}\, \beta_{i,m} + x^{M}_{i,m,\ell}{}' \gamma_{i,m}, \text{ and } \left(\beta'_{i,1:M,0}, \beta'_{i,1:M}, \gamma'_{i,1:M}\right)' \sim MVN\left(\overline{\beta}, D_{\beta}\right),$$

where $\alpha_{i,1:J} \equiv vec(\alpha_i)$, with α_i a $\left(J \times P^{A}\right)$ matrix with P^{A} the number of elements in $x^{A}_{i,j,\ell}$, $\gamma_{i,1:M} = vec(\gamma_i)$ with γ_i a $\left(M \times P^{M}\right)$ matrix with P^{M} the number of elements in $x^{M}_{i,m,\ell}$, and $\beta_{i,1:M,0}$ and $\beta_{i,1:M}$ are $(M \times 1)$ vectors. The parameters follow multivariate normal distributions, as shown in (2) and (3), to account for heterogeneity among consumers. We provide more details on the explanatory variables included in equations (2) and (3) next.

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. 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). These variables are included in $x^{A}_{i,j,\ell}$ in equation (2). Brand familiarity (as a person-related factor) is also predicted to influence attention to the ad and its elements (cf., Reichle, Rayner, and Pollatsek 2003). Therefore, this variable is included in $x^{A}_{i,j,\ell}$ in equation (2). The parameters $\alpha_{i,1:J}$ reflect the direct effects of these variables on attention.

The probability that a consumer claims to have noted the ad (m = 1), seen 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 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, $E[n_{i,j,\ell}] \cdot E[d_{i,j,\ell} \mid n_{i,j,\ell}] = \mu_{i,j,\ell} \cdot \lambda_{i,j,\ell}$. Clearly, the brand-associated measure depends on attention to the brand element ($m = j = 2$), and the read-most measure depends on attention to the text element ($m = j = 3$), and these are specified in equation (3) as $\phi_{i,m,\ell} = \mu_{i,m,\ell} \cdot \lambda_{i,m,\ell}$. The ad-noted measure depends on attention to the entire ad comprised of the three elements, so that $\phi_{i,1,\ell} = \sum_j \mu_{i,j,\ell} \cdot \lambda_{i,j,\ell}$ (for m =1) in equation (3). The parameters $\beta_{i,m}$ are the individual-specific attention weights, capturing the effects of attention on recognition memory. Recognition is claimed when a consumer-specific threshold, $-\beta_{i,m,0}$, 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 ad-specific 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 text element may lead consumers to believe that they must have read most of the ad and a large brand element may lead them to infer that they did not. 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). To allow for these effects, we include the size of the ad element and brand familiarity in $x^{M}_{i,m,\ell}$ in equation (3). The parameters $\gamma_{i,1:J}$ 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 allows for tests of the effects of ad layout and brand familiarity 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 effects would demonstrate systematic biases in the recognition scores, and reduce their diagnostic value.

Model Estimation

A Metropolis-within-Gibbs sampling algorithm is used to estimate the model, with standard diffuse priors for all distributions (Rossi, Allenby, and McCulloch 2005). Matrix-normal priors are used for all regression coefficients with mean zero and variance $10^4 I$, and for the variance-covariance matrices $D_{\cdot}$. 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. 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. 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[1].

DIAGNOSTIC VALUE OF AD RECOGNITION

The proposed model predicts recognition memory from prior attention to the ads and other factors. Deductively, it can be used to establish the probability that recognition is claimed when attention was actually devoted to the ad and its elements, and the probability that recognition is not claimed when attention was actually not devoted to the ad and its elements (see Altman and Bland 1994a). In advertising research and practice, however, ad recognition tests are used to make inferences about attention to ads during prior exposure in situations where the latter is not measured. In those applications one would like to know the accuracy of the recognition test as a diagnostic measure for attention, inductively. This is similar to medical decisions in which diagnostic tests are used to determine the true unknown status of people on a particular condition (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, 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 an individual has fixated on an ad or a specific ad element a certain number of times or more during exposure, given that s/he claims to have seen it, and define NDV as the probability that an

[1] The log marginal density computed with a harmonic mean across the draws of the MCMC chain is a popular choice in model comparisons, but is biased and invariably points to the largest model in the set.

individual has not fixated on an ad or a specific ad element less than 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:

$$\begin{aligned} \text{PDV}(\chi) &= p\left(n_{i,j,\ell} \geq \chi_{PDV} \mid I_{m,l} = 1\right) \\ \text{NDV}(\chi) &= p\left(n_{i,j,\ell} < \chi_{NDV} \mid I_{m,l} = 0\right). \end{aligned} \tag{4}$$

Equation (4) can be evaluated based on the parameter estimates obtained from the attention and memory model using Bayes theorem, as shown in Appendix I. 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 the diagnostic 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 (Leisenring and Sullivan Pepe 1998).

In summary, our current research extends earlier work on attention and memory as follows. First, the proposed model includes 1) a more accurate account of the process through which gaze arises, 2) a full heterogeneity specification of the effects of stimulus- and person-related factors on gaze and recognition, 3) multiple, correlated recognition measures, and 4) unobserved attention, rather than observed fixation frequencies, as a predictor of memory.

Second, the model allows for assessment of the indirect effects of ad layout and brand familiarity on ad recognition mediated by attention, and their direct effects over and above attention. These latter effects would be evidence of systematic biases in the recognition scores. 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 propose a bias-adjustment procedure for ad recognition measures based on this.

RESULTS

We compare the log-marginal density (LMD) of three nested alternative models to determine the contribution of specific factors to recognition memory, with a lower LMD indicating better fit. 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 2008). 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 -140505. The second model, which adds the direct effects of the brand, pictorial and text size on ad recognition, improves on this (LMD drops to -140041). 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 drops to -139600. 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 biased due to retrieval factors. We present parameter estimates of the third, full model.

Table 2 presents the parameter estimates for the attention part of the model first. 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. 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 ad recognition part of the model. There is clear evidence for attention effects on the ad-noted measure (the 95% posterior credible interval of the parameter not covering zero), and some evidence for the effect of brand attention on the brand-associated measure (the 90% posterior credible interval of the parameter not covering zero). 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 uniformly increases claimed ad, brand and text recognition, regardless of how much attention was devoted to the ad during the earlier exposure. These findings are consistent with research on the role of pictorial size on ad recognition measures (Finn 1988), but they show the effect to be independent of the actual

attention devoted to the pictorial. These results imply that ads with larger pictorials tend to attain higher brand-associated and read-most scores than warranted. Also, more text in the ad increases the probability of claiming recognition of the brand and text, regardless of how much attention was paid to these elements. The large positive direct effect of text-size on the read-most measure is particularly troublesome, because text-size does not influence attention indirectly, because attention to text does not influence text recognition. Conversely, larger brand sizes decrease the ad-noted and read-most measures, independent of the actual attention devoted to them during ad exposure.[2] Apparently, larger text and smaller brand elements serve as retrieval cues at the time of the recognition test that makes participants infer that more attention must have been devoted to the text during ad exposure. In addition, consumers over claim to have noted ads for familiar brands and to have read-most of their text, independent of their attention to the ads.

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

[2] 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.

threshold ($\chi = 0,1,2,...$). Appendix II provides the diagnostic value curves separately for each of the 48 ads in our study. 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, see, e.g., Charness, Reingold, Pomplun, and Stampe (2001), Masciochi, Mihalas, Parkhurst, and Niebur (2008), Torralba, Oliva, Castelhano, and Henderson (2006). It corresponds to roughly 1-2 seconds of exposure needed for reliable recognition memory, which reflects self-paced exposure durations to ads in natural conditions (Pieters and Wedel 2004).

*** Insert Figure 1 ***

The top panel of Figure 2 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 92.7%. 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 23.9% and 33.3%. To illustrate, at the threshold, the odds of the PDVs of ad-noted over brand-associated are almost 4:1 (0.93/0.24) 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 2 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 a high 81.8%. 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

only12.2%. 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 certainty (92.7%) 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 certainty (81.8%) that they fixated it less than 5 times. Although these diagnostic values are substantial they are not ideal, and the other diagnostic values of all three recognition measures are considerably lower.

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. 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 $\text{PDV}(\chi)$ and $\text{NDV}(\chi)$ that can be read directly from Figure 2 at the desired threshold $\chi = 5$ (or any other desired threshold), and is computed as follows:

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

Equation (6) is derived from the rule of total probability: $P(A) = P(A \mid B)P(B) + P(A \mid \overline{B})P(\overline{B})$. 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). $P(A \mid B)$ 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. 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 .92 and .13, respectively, then the BAR score is (.92)(.80) + (.87)(.20) = .91. In general, the BAR score can range from 0 to 1. When PDV and NDV sum to one, the bias-adjusted test 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.

Hold-out Validation. To demonstrate the improved accuracy of BAR scores over raw recognition scores, we re-estimate the model for a random sample of 38 ads, and retain 10 ads as a hold-out sample. We adjust the raw scores of the hold-out sample of ads using equation (6), with PDV and NDV estimated from the calibration sample, averaging PDV and NDV across ads for each of the recognition measures. We compute the absolute deviations of these BAR scores from the true scores $|BAR\ score_a - true\ score_a|$ and of the raw scores from the true scores $|raw\ score_a - true\ score_a|$ for each ad. 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. 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 (38 ads) and out-of-sample (10 ads) 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 $MAD_b < MAD_r$). In-sample MADs of the BAR scores are relatively small, 10.6%, 14.6% and 18.3%, respectively, for ad-noted, brand-associated and read-most. This is a substantial reduction from the MADs for the raw scores, which are around 25% (Table 4). Not surprisingly, the BAR score for read-most still performs worst, which 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, the out-of-sample MADs are even somewhat smaller for the brand-associated (8.9%) and read-most scores (13.3%). This may have been due to the specific ads in our (random) hold-out sample. We also compute the percentage improvement in bias-adjusted relative to raw recognition scores (Table 4). Improvement in accuracy ranges from roughly 10 to 25% out-of-sample, which is substantial.

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 beneficial in the context of natural exposure conditions, where attention to ads is short, in the order of magnitude of a few seconds (Pieters and Wedel 2004), and where ad recognition tests are frequently used. We therefore also apply the bias-adjustment procedure to the data collected after in-home exposure. Note that for the participants in the home condition eye-tracking data are not available. 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 from

the same population were randomly allocated to one of the two conditions, and the same set of ads was evaluated in the same editorial context. If common exposure and retrieval biases are present and removed by the bias-adjustment procedure, the scores between in-home and lab conditions should be closer after the correction.

In fact, the differences between the raw recognition scores in the in-home and lab conditions are substantial, 15.1% for ad-noted, 11.0% for brand-associated and 0.6% for read most. After bias-correction, these differences between conditions diminished substantially, to .7%, .6% and .1% respectively. These results are in part due to the low diagnosticity of the test and reveal that bias-adjustment brings the recognition scores of the in-home and lab conditions closer, and correct recognition scores collected after exposure in natural in-home settings, as they are frequently used in practice.

DISCUSSION

We found that attention during ad exposure predicted several ad recognition measures even after controlling for factors that may potentially bias them, and it partially mediated the effects of ad layout and brand familiarity on recognition. 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, the diagnosticity of ad recognition measures is far from ideal, and attention to the text in ads did not significantly affect the read-most measure. Systematic memory biases were apparent and there was also evidence for memory reconstruction based on the ad layout and brand familiarity. First, independent of attention, consumers over-claimed ad recognition when the ad contained a larger pictorial and a smaller 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. Second, independent of actual attention, consumers were more likely to report having noted an ad and read most of its text when the ad was for a familiar brand. Because brand familiarity also raised attention to the pictorial and text of ads, this revealed its double benefit: raising attention at exposure and raising subsequent recognition independent of attention. But, failure to control for brand familiarity in ad recognition measures may not only lead to overrating the quality of ads for familiar brands, but perhaps also their effectiveness in generating future sales, because familiar brands tend to have more frequent buyers (Ehrenberg, Goodhardt, and Barwise 1990).

The diagnostic value of ad recognition is low and varies across measures and metrics. Specifically, the positive diagnostic value of the ad-noted measure was high, but its negative diagnostic value was low. When consumers reported to have seen the ad, the probability that they looked at it for at least 1-2 seconds (five fixations), which is typical for natural exposure conditions, was over 90%, which is high. However, the likelihood of correctly reporting not having noted an ad when in reality consumers looked at it less than 1-2 seconds was only about 10%, which is very low. So, the ad-noted measure was better 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. When consumers claimed to recognize the brand and text in the ad the probabilities that they actually looked at them for 1-2 seconds were only 20-30%. Yet, the probabilities of correctly reporting not having noted the brand and text when in reality consumers actually did not look at them for 1-2 seconds were about 70-80%, which is

substantial. So, the brand-associated and read-most recognition measures were better at excluding ads that were actually not noted. None of the ad recognition measures performed well in both accurately identifying noted and excluding unnoted ads.

Our proposed procedure was based on a threshold of five fixations selected based on theory and prior research. However, it allows for sensitivity analyses of diagnosticity to threshold values (Altman and Bland 1994c) different from the threshold of five fixations. Such analyses 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 brand-associated there was no threshold where the positive and negative diagnostic values both exceeded .50. Taken together, these findings cast doubts on the diagnostic value of ad recognition measures for attention during prior ad exposure which they purport to reflect, and challenge their use in advertising theory and practice. However, as yet, it is not clear beyond which specific positive and negative diagnostic values ad recognition tests are still useful. This is an important topic for future research.

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 challenging (Puntoni and Tavassoli 2007; Whittlesea and Leboe 2000), because such factors may influence both exposure and retrieval, and in different ways. For instance, in our study the size of the text element

decreased attention to the brand, but increased brand-associated 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."[3] Thus, ad-noted, brand-associated and read-most scores across magazines and product categories benchmark the effectiveness of print advertising[4], and similar recognition measures are used in television advertising.[5] They are being used to assess ad effectiveness and to determine "which ads attract the most attention,"[6] and serve as inputs to advertising message and media decisions.[7] 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 familiar brands and 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 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 (Appendix II). One reason why ad recognition measures are recommended in advertising research is their presumed ability to detect delicate attentional and perceptual processes during

---

[3] http://www.mcnairingenuity.com.au, accessed July 2008.

[4] 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.

[5] www.ameritest.net/products/adtracking.pdf, accessed July 2008.

[6] http://www.time.com/time/mediakit/audience/research/proprietary/starch.html, accessed July 2008.

[7] 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, July 2008.

exposure (Heath and Nairn 2005). The present findings indicate that they may unfortunately have insufficient diagnostic value for this purpose.

Starch-type recognition measures have a long tradition in advertising practice, and are relatively easy and cheap to collect. Eliminating them may lead to undesirable regime-switches in measurement of ad effectiveness for a large number of companies relying on them. In those cases, the bias-adjustments that we have developed and tested may be gainfully used to remove biases from these recognition scores. The bias-adjustments may improve the accuracy of recognition scores for 1-2 seconds exposure durations by as much as 10-25%. In addition, the bias-adjustments go a long way in removing differences between tests conducted after in-home and lab exposure conditions.

Because we could not track consumers' eye-movements at home, we were not able to assess biases at exposure in 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 due to retrieval cues than what we observed in the lab condition, which future research may further investigate. Research is also called for to improve the accuracy of measurement instruments for ad-recognition tests. Focus should be on improving the low negative diagnostic value of the ad-noted score, and the low positive diagnostic value of the brand-associated and read-most scores. Triangulation with other memory measures, such as recall and indirect measures of memory appears to be one viable route to do so (Krishnan and Chakravarti 1999), but the joint measurement of attention and memory as proposed here seems most fruitful.

We developed a framework to assess and improve recognition test diagnosticity in marketing, comprising three interrelated components. First, it contains a model of the

relationship between attention during ad exposure and subsequent recognition memory. The model extends previous accounts of the relationship between attention and memory, by specifying both attention and memory as latent constructs, with attention reflected in both fixation frequency and duration, and recognition claimed when a memory threshold is exceeded. The model accommodates direct effects of covariates on attention and on memory, and their indirect effects on memory mediated by attention. This model formulation provides unbiased tests of the diagnostic value of memory measures for prior attention. Second, we propose to quantify positive and negative diagnostic values of ad recognition measures using the parameter estimates of the model, and obtain inverse probabilities that consumers have (not) actually attended a specific ad and its elements given that they do (not) claim recognition. The metrics of diagnosticity are estimated simultaneously with the process model, controlling for the potential effects of the covariates and accounting for uncertainty in parameter estimates. Third, we develop a procedure to arrive at bias-adjusted recognition (BAR) scores. The bias-adjustment procedure affords more accurate inferences about attention during ad exposure based on the raw recognition scores.

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. Other applications include short-form scales or single-items which are gaining prominence in marketing research as quick screeners when large multi-item scales are too intrusive or costly to implement, such as in web or telephone surveys. The positive and negative diagnosticity of the short-form or single items for the multi-item scales can be established and improved based on a model of the relationship between them. In a similar vein, the diagnosticity of recognition tests of outdoor, television and web advertising could be assessed and corrected.

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.

APPENDIX I:
DERIVATION OF INVERSE CONDITIONAL PROBABILITIES

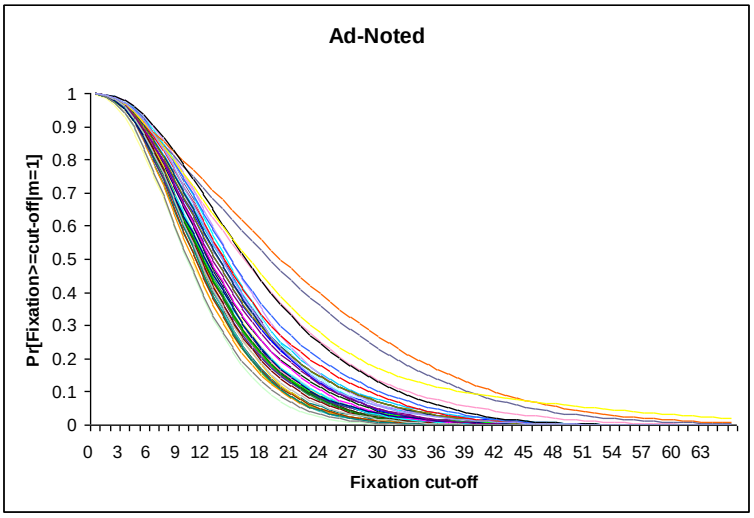
The inverse probability that an individual has fixated on the ad or ad element, in case s/he claims (no) recognition, $p\left(N_{i,j,l} = n_{i,j,l} \mid y_{i,m,l}\right)$ is computed as:

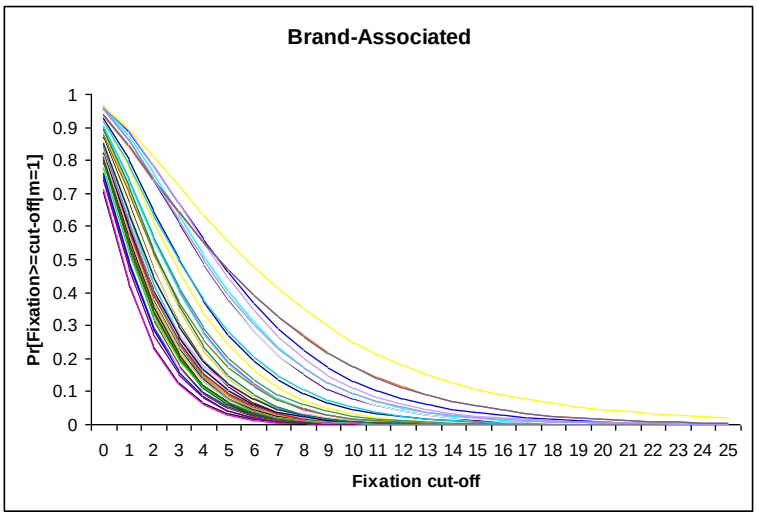
$$\text{(A1)} \qquad \frac{\int_{\omega}\int_{\lambda}\int_{\mu}\int_{t} f_N\left(n_{i,j,l} \mid \mu_{i,j,l}\right) f_{T|N}\left(t_{i,t,j} \mid n_{i,j,l}\right) f_Y\left(y_{i,m,l} \mid \mu_{i,j,l}\lambda_{i,j,l}, \omega_{i,m,l}\right) p(\mu_{i,j,l})\, p(\lambda_{i,j,l})\, p(\omega_{i,m,l})\, dt\, d\mu\, d\lambda\, d\omega}{\int_{\omega}\int_{\lambda}\int_{\mu} p\left(y_{i,m,l} \mid \mu_{i,j,l}\lambda_{i,j,l}, \omega_{i,m,l}\right) p(\mu_{i,j,l})\, p(\lambda_{i,j,l})\, p(\omega_{i,m,l})\, d\mu\, d\lambda\, d\omega}$$

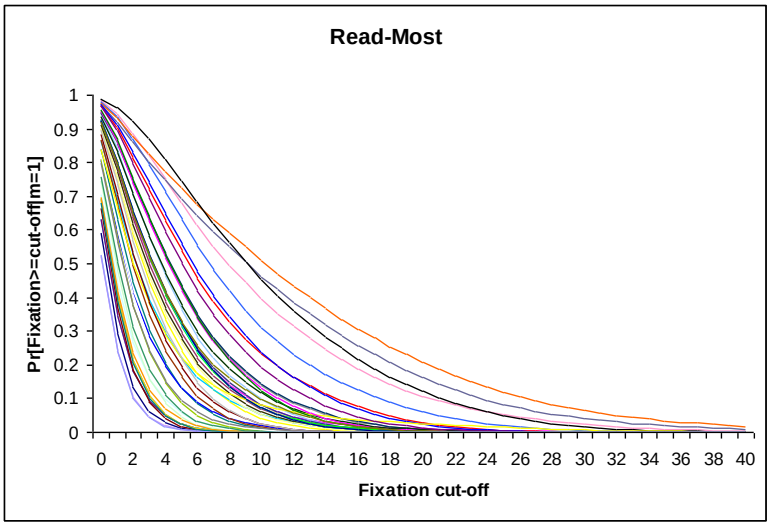
Here $f_N\left(n_{i,j,\ell} \mid \mu_{i,j,l}\right)$ is the conditional probability of observing $n_{i,j,l}$ fixations given $\mu_{i,j,l}$, and $f_{T|N}\left(T_{i,j,l} \mid n_{i,j,l}\right)$ is the conditional density of observing gaze duration $t_{i,j,l}$ given $n_{i,j,l}$ and $\lambda_{i,j,l}$ for ad element *j* associated with consumer *i* and ad *l*. $f_Y\left(y_{i,m,l} = 1 \mid \mu_{i,j,l}\lambda_{i,j,l}, \omega_{i,m,l}\right)$ 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 $\phi_{i,j,\ell} = \mu_{i,j,\ell}\lambda_{i,j,\ell}$ 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, $\phi_{i,j,\ell}$. Note that we use fixation frequency as the basis for computing the PDV and that the numerator in Equation (A1) 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 $f_Y(y_{i,m,l} = 1 \mid \mu_{i,j,l}\lambda_{i,j,l}, \omega_{i,m,l})$ based on Equation (1) $f_Y(y_{i,m,l} = 1 \mid \mu_{i,j,l}\lambda_{i,j,l}, \omega_{i,m,l})$ is used for the denominator of the PDV and $1 - f_Y(y_{i,m,l} = 1 \mid \mu_{i,j,l}\lambda_{i,j,l}, \omega_{i,m,l})$ for the denominator of the NDV. Next, we compute $(1 - f_N(n_{i,j,l} < \chi \mid \mu_{i,j,l})) \times f_Y(y_{i,m,l} = 1 \mid \mu_{i,j,l}\lambda_{i,j,l}, \omega_{i,m,l})$ for the numerator of the PDV and $f_N(n_{i,j,l} < \chi \mid \mu_{\mathrm{i,j,l}}) \times (1 - f_Y(y_{i,m,l} = 1 \mid \mu_{i,j,l}\lambda_{i,j,l}, \omega_{i,m,l}))$ 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.

APPENDIX IIa

POSITIVE DIAGNOSTIC VALUE CURVES FOR INDIVIDUAL ADS

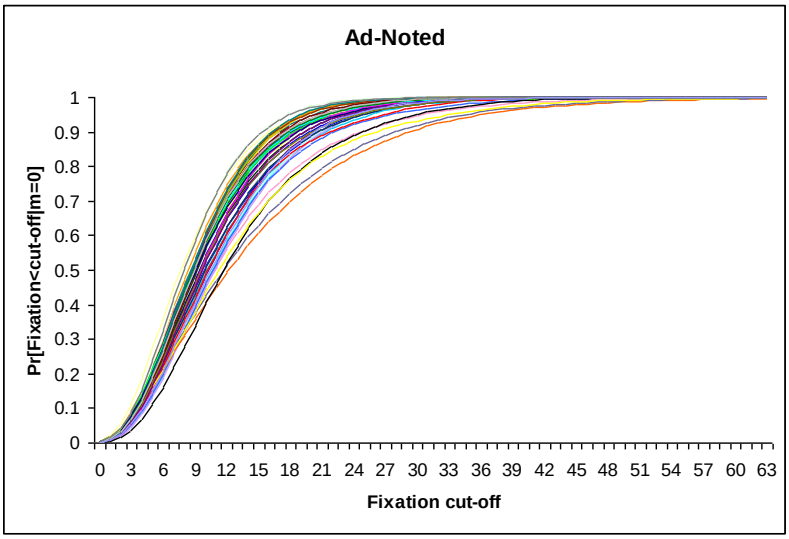


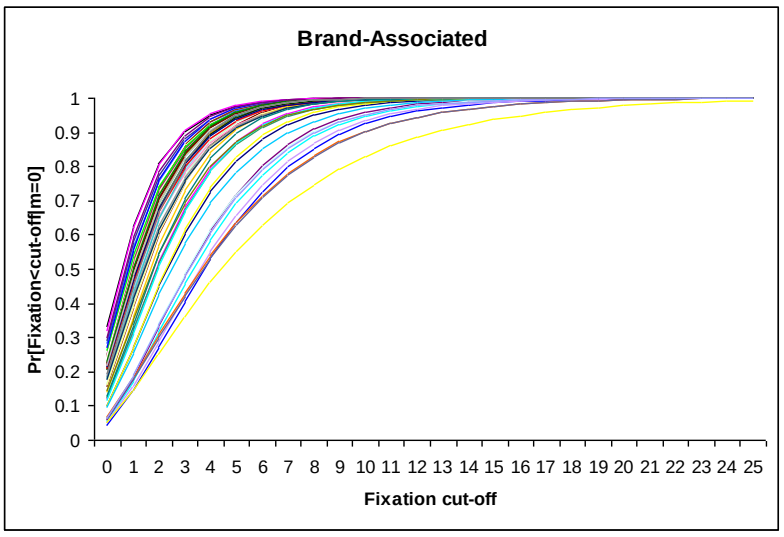


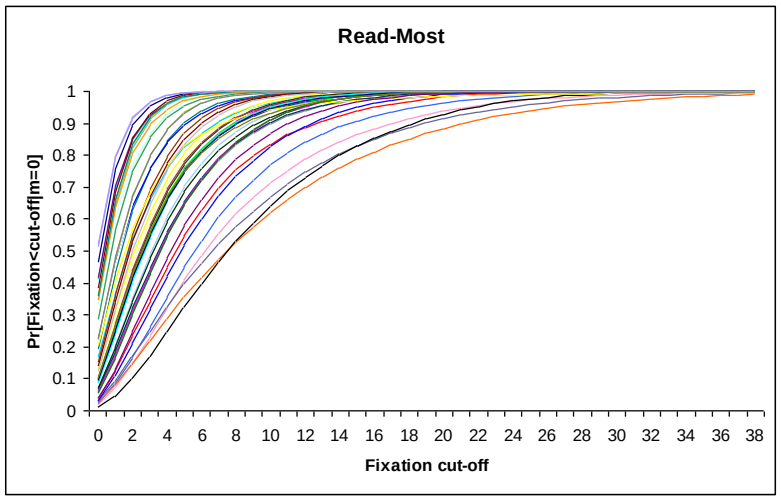


APPENDIX IIb

NEGATIVE DIAGNOSTIC VALUE CURVES FOR INDIVIDUAL ADS







REFERENCES

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

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

Altman, Douglas G. and J. Martin Bland (1994b), "Diagnostic Tests 2: Predictive Values," *British Medical Journal*, 309, 102.

Altman, Douglas G. and J. Martin Bland (1994c), "Diagnostic Tests 3: Receiver Operating Characteristic Plots," *British Medical Journal*, 309, 188.

Baldinger, Allan L. and William A. Cook (2006), "Ad Testing," in Rajeev Grover and Marco Vriens (eds.), *Handbook of Marketing Research* (pp. 487-505), London: Sage.

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

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,* 1146-1152.

Chib, Siddharta (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association,* 90, 1313-1321.

Chib, Siddharta and Ivan Jeliazkov (2001), "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association,* 96, 270-281.

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), 321-334.

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

Finn, Adam (1988), "Print Ad Recognition Readership Scores: An Information Processing Perspective," *Journal of Marketing Research*, 25, 168-177.

Finn, Adam (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), 1783-1792.

Hanssens, Dominique M. and Barton A. Weitz (1980), "The Effectiveness of Industrial Print Advertisements Across Product Categories," *Journal of Marketing Research*, 17, 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), 419-432.

Havlena, William J. and Jeffrey Graham (2004), "Decay Effects in On-line Advertising: Quantifying the Effect of Time Since Last Exposure on Branding Effectiveness," *Journal of Advertising Research*, 44, 327-332.

Heath, Robert and Agnes Nairn (2005), "Measuring Affective Advertising: Implications of Low Attention Processing on Recall," *Journal of Advertising Research*, 45 (2), 269-281.

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, 281-292.

Henderson, John M. (1992), "Object Identification in Context: The Visual Processing of Natural Scenes," *Canadian Journal of Psychology*, 46 (3), 319-341.

Hintzman, Douglas, L. (2000), "Memory Judgments, " in Endel Tulving and Fergus I. M. Craik (eds.), *The Oxford Handbook of Memory* (pp. 165-178), Oxford, UK: The Oxford University Press.

Itti, Laurent (2005), "Models of Bottom-Up Attention and Saliency" in Laurent Itti, Geraint Rees, and John K. Tsotsos (eds.), *Neurobiology of Attention* (pp. 576-582), Amsterdam: Elsevier Academic Press.

Janiszewski, Chris (1998), "The Influence of Display Characteristics on Visual Exploratory Search Behavior," *Journal of Consumer Research*, 25 (December), 290-301.

Johnson, Marcia K., Shahin Hashtroudi, and D. Stephen Lindsay (1993), "Source Monitoring," *Psychological Bulletin*, 114, 3-28.

Kelly, Colleen M and Larry L. Jacoby (2000), "Recollection and Familiarity," in Endel Tulving and Fergus I.M. Craik (eds.), *The Oxford Handbook of Memory* (pp. 215-228), Oxford, UK: Oxford University Press.

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, 444-452.

MacKinnon, David P., Amanda J. Fairchild, and Matthew S. Fritz (2007), "Mediation Analysis," *Annual*

Review of Psychology, 58, 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.

Mitchell, Karen J. and Marcia K. Johnson (2000), "Source Monitoring," in Endel Tulving and Fergus I. M. Craik (eds.), *The Oxford Handbook of Memory* (pp. 179-195), Oxford, UK: Oxford University Press.

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, 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, 141-148.

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.

Pieters, Rik and Michel Wedel (2007), "Informativeness of Eye Movements for Visual Marketing: Six Cornerstones", in Michel Wedel and Rik Pieters (eds.), "*Visual Marketing: From Attention to Action*," (pp. 43-71), New York: Lawrence Erlbaum, Taylor & Francis.

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

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, 445-526.

Roediger, Henry L. and Kathleen B. McDermott (2000), "Distortions of Memory," in Endel Tulving and Fergus I.M. Craik (eds.), *The Oxford Handbook of Memory* (pp. 149-162), Oxford, UK: Oxford University Press.

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), 118-124.

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), 327-336

Singh Surendra N., 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.

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), 766-786.

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) 219-232.

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

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, 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*, 441-517.

Zhang, Jie, Michel Wedel, and Rik Pieters (2008), "Sales Effects of Feature Advertisements: A Bayesian Mediation Analysis," *Journal of Marketing Research*, forthcoming.

TABLE 1
DESCRIPTIVE STATISTICS

Variable	N	Mean	SD	Median	Min	Max
Ads: Surface sizes:						
Brand (dm^2)	48	0.732	0.555	0.537	0.120	3.008
Pictorial (dm^2)	48	4.270	1.067	4.506	0.578	5.367
Text (dm^2)	48	1.019	0.761	1.104	0.000	3.086
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	**.776**	.047	**1.641**	.034	**.900**	.051
Surface size:						
Brand	**1.230**	.050	**.132**	.040	**-.769**	.067
Pictorial	**-.588**	.055	**.858**	.051	*-.079*	.053
Text	**-.543**	.051	**-.208**	.037	**1.751**	.044
Brand familiarity	.024	.027	**.049**	.025	**.102**	.028
Covariances for fixation frequency:						
Brand	**.399**	.043	(.582)		(.736)	
Pictorial	**.174**	.027	**.226**	.025	(.544)	
Text	**.333**	.042	**.185**	.029	**.516**	.056
Fixation duration:						
Ln(Mean)	**-.800**	.026	**-.346**	.029	**-.634**	.030
Ln(Dispersion)	**-1.462**	.013	**-1.821**	.015	**-1.336**	.015

Note - Bolded (italicized) parameter estimates indicate that probabilities of the parameters to be larger or smaller than zero are greater than .95 (.90). Correlations are between parentheses. Dispersion is equal to variance divided by mean squared for a random stopped sum model with fixation frequency distributed as Poisson and fixation duration as gamma.

TABLE 3
DETERMINANTS OF AD RECOGNITION

	Ad recognition memory					
	Ad noted		Brand associated		Read most	
Predictors	Mean	SD	Mean	SD	Mean	SD
Intercept	**-.601**	.096	**-.742**	.065	**-.717**	.057
Latent attention	**.104**	.012	*.047*	.032	.014	.019
Surface size:						
Brand	**-.420**	.077	.032	.081	**-.332**	.069
Pictorial	**.320**	.097	**.272**	.092	**.412**	.089
Text	.027	.065	**.178**	.068	**.351**	.077
Brand familiarity	**.145**	.029	-.015	.022	**.068**	.024

Note - Bolded (italicized) parameter estimates indicate that probabilities of the parameters to be larger or smaller than zero are greater than .95 (.90). Correlation between brand-associated and read-most measures is .158 (SD = .020)

TABLE 4
BIAS-ADJUSTMENT OF RECOGNITION MEMORY

	Ad Noted (%)	Brand Associated (%)	Read Most (%)
PDV	92.7	23.9	33.3
NDV	12.2	81.8	71.8
In-sample:			
MAD raw score (MAD_r)	27.7	24.1	24.1
MAD bias-adjusted score (MAD_b)	10.3	14.2	18.3
MAD_b-MAD_r	17.1	9.9	5.8
% improvement	23.8	20.4	13.4
Out-of-sample:			
MAD raw score (MAD_r)	21.7	23.6	21.8
MAD bias-adjusted score (MAD_b)	10.3	8.9	13.3
MAD_b-MAD_r	11.1	14.7	8.5
% improvement	14.1	25.9	15.1

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)