2-1994

# Signal Detection Theory and Single Observation Designs: Methods and Indices for Advertising Recognition Testing

Dennis J. Cradit
*Florida State University*

Armen Tashchian
*Kennesaw State University*, atashchi@kennesaw.edu

Charles F. Hofacker
*Florida State University*

J. DENNIS CRADIT, ARMEN TASHCHIAN, and CHARLES F. HOFACKER*

Most marketing applications of signal detection theory (SDT) produce an estimate of the respondent's memory accuracy based on exposure to a number of advertisements. Marketing practitioners, however, are usually more interested in the performance of an individual advertisement, or elements of that ad. Moreover, advertising recognition paradigms are typically limited to single observations per respondent. The authors present and compare two alternative methodologies that estimate SDT parameters for such designs by pooling recognition performance across respondents. They present two simulations that explore the most efficient methodology and suggest guidelines for selecting appropriate accuracy indices.

# Signal Detection Theory and Single Observation Designs: Methods and Indices for Advertising Recognition Testing

In one of the earliest examples of a memory-recognition measure of ad effectiveness, Lucas (1942) showed consumers an advertisement to which they had been exposed previously and asked if they could remember having seen the ad when they read an issue of the magazine in which the ad originally appeared. He then averaged these consumer responses to produce an estimate of the percentage of the sample reporting recognition of the ad; producing, in effect, a percentage-readership score. The underlying assumption of Lucas' (1942) procedure, as well as of more recent measures of recognition (Bagozzi and Silk 1983; Singh and Rothschild 1983), is that an assessment of the advertisement's "familiarity" or "memorability" constitutes an indication of the ad's relative effectiveness.

The strength of a particular recognition methodology is evidenced by its ability to adjust for the presence of respondent bias. Test participants inflate memory performance by reporting recognition of advertisements that they, in fact, have never seen. Appel and Blum (1961) included distractor-advertisement trials in a recognition test (i.e., advertisements constructed solely for purposes of testing and that respondents *ipso facto* cannot remember), and found that a significant percentage of respondents report recognition of these fictitious ads. In fact, an advertisement's percentage-readership score is composed of a true-recognition component reflecting overall memorability of the advertisement across the consumer sample, and a bias component that inflates or deflates the score, depending on the number and severity of yeasayers and naysayers in the sample (Wells 1961). The challenge for advertising research is to develop a method that will separate these two components. Typical approaches to this problem use a mixture of distractor (fictitious) and target (real) advertisement trials to produce estimates of hit rates (correct detection of target advertisements) and false-alarm rates (incorrect recognition of distractor ads), which in turn can be used as inputs for the computation of indices that attempt to produce some measure of actual memorability, corrected for yeasaying and naysaying response biases (Singh and Churchill 1987; Swets and Pickett 1982).

Recently, a number of papers in marketing and advertising suggest the use of an elegant mathematical model known as signal detection theory (SDT) to improve ad recognition testing. SDT is used to analyze experimentally produced hit and false-alarm rates within a decision-theoretic framework to produce a consistent and reliable estimate of the respondent's actual memory

*J. Dennis Cradit (formerly J. Dennis White) and Charles F. Hofacker are Associate Professors of Marketing, College of Business, Florida State University. Armen Tashchian is a Professor of Marketing, Department of Marketing, Kennesaw State College. The authors appreciate the many valuable changes suggested by the *JMR* reviewer.



*Journal of Marketing Research*
Vol. XXXI (February 1994), 117–27

accuracy, and a separate estimate of the respondent's tendency to over- or under-report recognition of the target stimulus (their "decisional criterion," in SDT terminology) (Banks 1970; Green and Swets 1966; Swets and Pickett 1982). SDT assumes that with the presentation of each ad in a recognition test, the respondent experiences some subjective sense of familiarity (denoted as a real-valued number, $X$) and uses this feeling of familiarity as evidence to determine if the test ad is a target or a distractor. Target ads and distractor ads generate overlapping distributions of evidence with the mean of the target ad distribution logically higher on the familiarity continuum than the mean of the distractor ad distribution. This reflects the fact that though the majority of target ads will result in feelings of familiarity stronger than those elicited by the majority of distractor ads, there are cases in which a distractor ad can elicit a strong feeling of familiarity whereas a target ad might elicit only a weak feeling of familiarity. SDT further assumes that the respondent sets a decisional criterion or threshold, $X_c$, on the familiarity continuum. On each trial, the respondent compares the location of the $X$ with the location of the decisional criterion. If feelings of familiarity exceed the criterion ($X > X_c$), the respondent reports recognition of the ad; if the feelings fail to exceed the criterion ($X < X_c$), the respondent reports no recognition of the ad. Such a situation enables the advertising researcher to represent recognition performance in two different ways. The respondent's performance can be summarized in a single index. One such index is the traditional $d'$ statistic (formally presented in equation 6), defined as the distance between the means of the two distributions when both are assumed normal. This distance, as we demonstrate, can be deduced from hit and false-alarm rates computed from the respondent's recognition data. In addition, recognition performance can be presented graphically through the use of the receiver operating characteristic (ROC) curve. This curve represents the respondent's trade-off between hits and false-alarms across different levels of decisional criteria, ranging from liberal (yeasaying) to conservative criteria (naysaying). Recognition performance is then indicated by computing the area under the ROC curve (see Tashchian, White, and Pak 1988 for a discussion).

However, though SDT represents a potential benefit to the area, its application to advertising research raises an important methodological problem. SDT traditionally has been applied in experimental psychology primarily to estimate memory accuracy of individual subjects by examining their performance across a range of experimental stimuli. The memorability of a specific stimulus within this range is rarely of consequence. In contrast, marketing practitioners, ordinarily unconcerned with the accuracy of individual respondents, need a bias-free estimate of the memorability of particular ads, or components of those ads. In effect, experimental psychologists seek to study individuals by aggregating discrimination behavior across stimuli; advertising researchers study stimuli, aggregating across individuals.

This difference in goals is aggravated by differences in available data. Requisite hit and false-alarm rates necessary to traditional SDT procedures require multiple observations per subject. In typical memory applications this rarely presents a problem because subject performance can be assessed over multiple presentations of stimuli. Ad testing methodologies, by contrast, must rely on a single recognition response per subject.

To date, two alternative aggregation approaches have been proposed to solve the problem of single observations per subject. Singh and Churchill (1986, 1987) suggest that respondents' recognition abilities, derived by observing their performance across a range of ads (one of which is the target ad of interest), be corrected for each respondent's level of bias (based on SDT-supplied indices). The average of these adjustments then can be computed across the sample and subtracted from the sample's hit rate, producing what is, in effect, a "bias-adjusted" version of Lucas' (1942) percentage-readership score. In contrast, Macmillan and Kaplan (1985) suggest a procedure that, when applied to ad testing, involves collapsing recognition performance for one particular ad across all respondents within a sample to produce a group estimate of memory accuracy for that particular ad. Though Singh and Churchill (1986) have provided preliminary reports of reliability for their procedure, little is known about its level of statistical bias or efficiency. Also, though the collapsed-index procedure has been suggested and discussed in marketing (Leigh and Menon 1986), no applications to single-observation designs or tests of its validity have yet been reported.

The single-observation design poses a related issue that must be addressed in transferring SDT to ad testing: that of the particular choice of accuracy estimate. Within the psychological literature, a number of sensitivity indices have been proposed for detection tasks, ranging from parametric measures based on Gaussian and logistic distributions to several nonparametric indices (Swets 1986a). When single-observation designs prompt the use of collapsing procedures, the selection of the index becomes particularly important because such a collapsing procedure requires the researchers to presume constant decision rules and constant sensory decision axes across all subjects in the sample (Macmillan and Kaplan 1985). Researchers might be hesitant, therefore, to assume the existence of normal distributions and response-criteria homogeneity. To date, most papers in marketing focus on the use of nonparametric indices, implicitly assuming the superiority of these over Gaussian-based measures. We review recent evidence (Swets 1986a, 1986b) that disputes this assumption and discuss implications of this evidence for choosing appropriate indices for a collapsed index of memory accuracy.

Our objective is therefore twofold. First, we test the relative performance of Singh and Churchill's (1986) adjustment procedure against Macmillan and Kaplan's

(1985) collapsing procedure. We present each methodology, describing the specific data-collection assumptions, and then test the relative validity of each approach in a simulation that assumes a wide range of true ad familiarity and decisional bias. Second, we evaluate the performance of nonparametric indices within a collapsing procedure, comparing them with more traditional measures based on Gaussian distributions. The results present a useful review of the relative advantages and limitations of the various indices for the assessment of memory for ads based on single observations per subject.

## *TWO ALTERNATIVE APPROACHES TO SINGLE-OBSERVATION DESIGNS*

In examining SDT procedures, one must observe the distinction between the method by which the hits and false alarms are collected and the particular computational formulas applied to those data. SDT data can be collected under several alternative procedures (see Tashchian, White, and Pak 1988 for a review) and analyzed by a number of alternative measures of memory accuracy. Though certain data collection methods presume a particular accuracy estimate, in most cases the researcher has a choice. Though both the Singh and Churchill (1986) and Macmillan and Kaplan (1985) approaches rely on standard SDT data collection procedures and computational formulas, each differs in the manner in which the data are aggregated and the sequence of analyses. In particular, the Singh and Churchill approach relies on a nonstandard application of a traditional measure of memory accuracy.

### *The Bias-Adjustment Approach*

Singh and Churchill's (1986, 1987) bias-adjustment approach is a variation of the traditional corrected hit probability formula (Green and Swets 1966), which attempts to adjust the raw hit rate by use of some measure of response bias. In the most typical case, a subject's false-alarm rate is subtracted from the hit rate

$$H_c = h - f, \tag{1}$$

where $h$ refers to the subject's hit rate and $f$ is the false-alarm rate. This procedure uses the raw hit rate as the measure of memory performance, but adjusts it for response bias reflected in the subject's tendency toward false alarms. A common variation on equation 1 attempts to normalize the corrected values:

$$H_c' = \frac{(h - f)}{(1 - f)}. \tag{2}$$

Equations 1 and 2 both represent relatively intuitive corrections for guessing and have been employed in a number of psychological studies of recognition memory (e.g., Fisk and Schneider 1984).

Singh and Churchill's approach relies on the corrected hit probability concept, though they employ a correction factor that is more complicated than the simple false-alarm rate. They suggest respondents be provided with a portfolio of real ads (one of which is the target ad of interest) and distractor ads. Though the response to the target ad is of primary interest, responses to the remaining ads in the portfolio are used to produce $h$ and $f$, necessary for computation of the SDT measure of each respondent's decisional criterion. The process through which this measure is produced starts with the computation of

$$B_j = \sum_{i=1}^{N} B_i x_{ij}, \tag{3}$$

where $x_{ij}$ is a dummy variable coded 1 if subject $i$ reports recognition of advertisement $j$ and 0 otherwise, $N$ is the number of respondents, and $B_i$ is a measure of response bias for individual $i$ modified from Hodos (1970, see also Grier 1971),

$$B_i = \begin{cases} 1 - \dfrac{f_i(1 - f_i)}{h_i(1 - h_i)} & \text{if} \quad h_i + f_i \le 1 \\ \dfrac{h_i(1 - h_i)}{f_i(1 - f_i)} - 1 & \text{if} \quad h_i + f_i > 1. \end{cases} \tag{4}$$

Here, $f_i$ is the false-alarm rate and $h_i$ is the hit rate for subject $i$. Note that though $B_i$ is based on an individual's response to all ads within a test portfolio, it is used to compute an average adjustment index per ad, $B_j$.

This adjustment then is subtracted from the group hit rate for the target ad

$$H_c'' = h_j - B_j, \tag{5}$$

where $h_j$ is the group hit rate for ad $j$ and $H_C''$ is the final corrected hit probability.

Though equation 5 is similar in form to equations 1 and 2, it employs a correction factor drawn from a Gaussian model of signal detection, and, as such, requires assumptions unusual for a traditional corrected hit probability. As we discuss subsequently, it also is important to note that corrected hit probability formulas generally imply the existence of high-threshold models of memory and cognition that, in turn, predict theoretical ROCs that are "non-regular" in form and almost always at odds with empirical ROCs collected from actual recognition data (Swets 1986a, 1986b).[1] However, because of the nature

[1]Every sensitivity index implies a particular theoretical or predicted ROC, which is derived by solving the index formula for $h$ and then plotting hits as a function of false alarms for each level of the index. A regular ROC is defined as one that obeys the following: $f = 0$ only when $h = 0$ and $h = 1$ only when $f = 1$. In other words, the curve is interior to the unit-square ROC except at the extremes ($f = 1$ and $h = 1$, or $f = 0$ and $h = 0$). In contrast, non-regular ROCs permit points having $h > 0$ for $f = 0$ or $h = 1$ for $f < 1$. Empirical evidence collected across a wide variety of tasks and designs consistently produce regular ROCs (see Swets 1986a for a discussion). Therefore, a critical test of the validity of a potential index is the degree to which it predicts a regular ROC.

of Singh and Churchill's correction factor, it is not immediately apparent what model is implied by equation 5 or what the nature of the form of its theoretical ROC is. The Appendix shows that the theoretical ROC implied by equation 5 is difficult to predict. Because of this, the efficiency and bias of equation 5 will be evaluated through simulation.

*A Collapsed-Index Procedure*

As mentioned previously, SDT models require a relatively large number of trials per subject to ensure stable estimators. As Macmillan and Kaplan (1985) note, however, researchers frequently confront the need to apply SDT models to designs that produce few responses per subject. One obvious solution is to increase the number of usable observations by combining data across subjects. In such a situation, one can compute indices for each member of the group, despite the insufficient number of observations, and then average these across the sample (a process Macmillan and Kaplan call "averaging"), or one can derive aggregate hits and false alarms across the sample and then compute the index on these group proportions (a process they call "collapsing").

Macmillan and Kaplan's comparison of averaging and collapsing procedures reveals that if subjects differ in bias but not accuracy, the $d'$ computed from group proportions will generally be lower than the average of the individual $d'$s. This loss in true $d'$ is significant only if the range in bias scores from the sample is in excess of 1.5 standard deviations. In addition, if subjects differ in accuracy but not in bias, the resulting $d'$ based on collapsed data will generally be lower than the average $d'$ of the separate individuals. Again, the decrement is severe only if the original $d'$s differ by 1.5 or greater. Finally, values of $d'$ computed from collapsed-group proportions are always less variable than the average of the individual $d'$s. Moreover, this variability decreases as the discrepancy between the subjects increases. Macmillan and Kaplan's conclusion is that the computation of a collapsed $d'$ from averaged proportions produces reliable, relatively unbiased estimates of accuracy.

*Collapsing procedures for ad testing*. Collapsing procedures thus far reported in the literature assume that, at a minimum, a sufficient number of observations are available to at least provide an $h$ and $f$ for each subject. In other words, they assume multiple observations per subject, though far fewer than would generally be considered appropriate. Data collection in ad testing situations, however, typically produces only a single observation per subject. In such a setting, there are insufficient data to compute basic proportions for each individual.

The use of a collapsing procedure for single-observation designs therefore will require a modified format. Responses to the target ad would need to be combined with responses to selected distractor ads to provide the necessary hits and false-alarm rates. A procedure to accomplish this might involve a standard SDT confidence-rating technique (Banks 1970; Swets and Pickett 1982) in which respondents are shown an ad and asked to report their "confidence" in their memory for that ad along a $k$-point scale, in which the anchors are "Certain I did not see it" to "Certain I did see it."[2] A traditional SDT analysis of each respondent's recognition data would produce an estimate of that respondent's memory sensitivity ability aggregated across the portfolio of target ads. To estimate the familiarity of a particular target ad, we simply pool the recognition data for the target ad and the distractor ads across all respondents in the sample. For example, in a portfolio containing ten target ads and ten distractor ads, each respondent contributes 11 responses to the analysis of any particular target—his or her confidence rating for the target ad of interest and confidence ratings for each of the ten distractor ads. In effect, subject responses correspond to trials within the more traditional SDT analysis. These raw responses are then decomposed into $k - 1$ hit and false-alarm pairs, which become $k - 1$ ROC points according to standard SDT confidence-rating procedures (see Tashchian, White, and Pak 1988 for an example).

*Suitable indices of performance*. A number of indices can be computed from the collapsed data, based either on a single $h$ and $f$ pair or, as in the case of confidence ratings, multiple pairs. The most obvious choice would be the traditional $d'$ statistic. The computational formula for the single-pair case would be

$$d' = Z_{h_j} - Z_f. \tag{6}$$

Here, $Z_{h_j}$ is the $z$-score associated with the hit rate for ad $j$, and $Z_f$ is the $z$-score associated with the overall false-alarm rate. In the multiple-pair case, $d'$ is estimated through an iterative estimation technique such as maximum likelihood (Dorfman and Alf 1969). In either setting, this statistic is based on the assumption of normal distributions of signal and noise with equal variance.

Reluctance to make such assumptions has led several researchers to suggest nonparametric indices, the most

[2]Normally, these $k$ response categories are used to compute $(k - 1)$ points on an ROC in the following manner: Assume that those responses to stimulus ads falling in the highest confidence category result from the respondent's strictest decision criterion. This is comparable to a yes/no task in which the respondent is induced to adopt a very conservative decision criterion. The responses in this highest confidence category are counted as "Yeses," and the responses in the remaining confidence categories are all counted as "Nos," and a hit/false-alarm pair is constructed representing the strictest decision criterion. Similarly, assume that those responses falling in the next highest confidence category result from the respondent setting a slightly less stringent decision criterion. Now, the responses in both the first and second highest confidence categories are counted as "Yeses," the responses in the remaining $k - 2$ categories are counted as "Nos," and a second hit/false-alarm pair is constructed representing the second strictest decision criterion. This process is repeated, cumulatively, across all $k$ categories of the response scale, resulting in $(k - 1)$ $2 \times 2$ conditional-probability matrices, and hence, $k - 1$ ROC points. Details and rationale for this procedure are available from several sources (Banks 1970; McNicol 1972; Tashchian, White, and Pak 1988).

popular of which in marketing is $A'$, which estimates the area under the ROC with only a single pair of hit and false-alarm rates:

$$A' = \frac{1}{2} + \frac{(h_j - f)(1 + h_j - f)}{4 - h_j(1 - f)}. \quad (7)$$

Here, $h_j$ is the aggregate hit rate for ad $j$, or ad component $j$, and $f$ is the overall false-alarm rate for the study. This measure runs from 0 to 1 and, as is well known, can be interpreted as percentage correct in a two-alternative forced-choice methodology (see Green and Swets 1966).

Swets (1986a, 1986b) reports extensive evidence that consistently shows that empirical ROCs are fitted well on a binormal graph by straight lines of varying slope. In other words, empirical ROCs are regular in shape and require a free slope parameter to adequately fit the data. Because $d'$ implies binormal ROCs that are linear with a fixed slope = 1.0, it would seem to be a poor choice for a suitable index (Swets 1986a).

As a result, Swets argues that the most appropriate index is $A_z$, the area under an ROC (on ordinary probability scales) that is consistent with empirical ROCs (Swets 1986b). Though estimated iteratively (Dorfman and Alf 1969) in the multiple hit and false-alarm case, when there is a single $h$ and $f$ pair, the index can be defined simply as

$$A_z = \Phi(d'/\sqrt{2}), \quad (8)$$

where $\Phi$ is the normal distribution. $A_z$ does not assume normal distributions, but any form of distribution that can be monotonically transformed to the normal (Swets 1986a). In contrast to $A'$, which is susceptible to poor placement of the single $h$ and $f$ pair along the ROC, $A_z$ can be calculated by fitting a straight line to multiple data points (plotted on a binormal graph) and is the more efficient, robust measure.

### *Summary*

Researchers faced with single observation designs have two general methodologies available for computing SDT measures of memory sensitivity: a bias-adjustment approach producing $H''_C$ and a collapsing procedure that produces $A_z$ or $A'$. The immediate issue is how to evaluate the two approaches. We can compare $A_z$ and $A'$ by examining the underlying models of memory and cognition that each implies. As mentioned previously, this is accomplished by determining the theoretical form of the ROC predicted by each particular index and then comparing this with empirical ROCs produced from research. Past work using $A_z$ shows that its theoretical and empirical ROCs do indeed coincide (Swets 1986a). Collapsing procedures utilizing $A_z$ therefore would seem to possess face validity. In contrast, Macmillan and Kaplan (1985) show that the ROC implied by $A'$ cannot be represented by a straight line when plotted on a binormal graph, a condition obviously at odds with empirical research. Use of this statistic with collapsing procedures should be approached with caution, though as we discuss subsequently, the computational simplicity of $A'$ nevertheless could lead us to overlook this theoretical deficiency.

As discussed previously, the theoretical predictions of $H''_C$ are more ambiguous. The traditional formulas presented in equations 1 and 2 imply various forms of a threshold model of memory recognition, at odds with empirical research (Swets 1986a). Because $H''_C$ is a complicated variation on the traditional formula, it also should predict theoretical ROCs at odds with empirical research. However, the bias-adjustment approach described by Singh and Churchill (1986) combines group-level hit rates with individual-level measures of bias. As such, it is difficult to determine a theoretical ROC from such an equation without several additional assumptions (e.g., the specific distributions of hits and false alarms across subjects, the sample size, the number of ads, and hit rates for other ads; see Appendix for a discussion). The result is that a direct comparison between $H''_C$ and the collapsed measures will require a simulation of their respective behavior.

## *TWO SIMULATIONS*

To compare the relative effectiveness of these two methodologies and the effects of collapsing procedures on the various accuracy indices, we conducted two computer simulations. Specifically, we sought to (a) compare statistical bias inherent in $H''_C$ with the area measures computed from collapsing procedures (i.e., determine if each familiarity estimate is, on average, equal to the true level of familiarity) and (b) measure the relative efficiency and consistency of estimates derived from the collapsed measures.

### *Simulation 1: Statistical Bias of the Estimator*

The first simulation directly compared $H''_C$, collapsed $d'$, collapsed $A'$, and collapsed $A_z$ to determine if each estimator was statistically unbiased. As a baseline comparison, performance of the alternative indices were compared with raw recognition scores (unadjusted hit rates). Because $H''_C$ and $A'$ are computed from single pairs of $h$ and $f$, a direct comparison required that we compute $A_z$ and $d'$ based on single pairs.

*Method.* The simulated data were generated assuming that each subject saw 48 ads and then was confronted with a test portfolio containing the original 48 ads and 48 new ads to which the simulated subject had not been exposed. The distractor ads were assumed normally distributed along the familiarity continuum with mean zero and unit variance. The 48 original ads were assigned familiarity means that increased from .0625 to 3.0 in 47 increments of .0625. The variance for each of the original ads was unity.

Four groups of 250 simulated subjects were constructed, each group employing a different decisional criterion. The criteria along the familiarity continuum for

reporting recognition of the ad were set at −.4, 0, .4, and .8. Therefore, the group with the −.4 criterion was most biased toward yeasaying whereas the group with the .8 criterion would be most biased towards naysaying. Subjects within each group were assumed to act on a personal decisional criterion, the variance of which was set at .2.

*Results*. The various panels in Figure 1 present the results of the simulation. Each panel displays the memory performance of the 48 ads, each ad displayed according to the four decisional criterion groups and plotted as a function of its respective true ad familiarity.

As a baseline comparison, Panel A shows the performance of the 48 ads in the four decisional criterion groups as measured by their uncorrected hit rates. As can be seen, the values rise in a clear monotonic manner as a function of the ads' true familiarity levels, as would be expected from the definition of hit rates. However, the inadequacies inherent in the index can be seen from the clear separation of ads by decisional criteria. The most liberal decisional criterion group (−.4), those respondents that would display the greatest amount of yeasaying, appear at the top of the graph, and the more conservative criterion groups are arranged in a predictable fashion below. In particular, note the many instances in which ads with the same hit rate originate from different levels of underlying familiarity.

A much different situation is presented in Panel B, which presents $A'$. This measure does a reasonable job of purging the various criterion groups of their decisional bias. The measure does show signs of statistical bias, however, when the true level of ad memorability exceeds 1.0. At mean levels above 1.0, the four groups begin to separate, thus implying that $A'$ confounds decisional bias and memory accuracy, though in a manner opposite to the raw, uncorrected hit rate. Essentially, $A'$ over-corrects the raw hit rate. As we discuss subsequently, marketing researchers should exercise caution when applying $A'$ in situations in which strong levels of familiarity are likely. Panel C shows the simulation results for $d'$. The four different response-bias groups are not consistently separated, which implies that $d'$ measures ad memorability unconfounded with response bias. At higher levels of ad memorability $d'$ becomes more variable, though in comparison with $A'$, $d'$ exhibits much less variability, and this variability is not systematically related to yeasaying/naysaying as is the case with $A'$. Note that $d'$ is not defined when hit rate is 1.0 and thus must be treated as missing data. As a result, much of the increasing variability of $d'$ displayed in Panel C is a direct result of simulated performance occasionally becoming perfect. Similar problems are evident in the pattern of data produced by $H''_C$, shown in Panel D. As with $A'$, $H''_C$ loses its ability to separate decisional bias and memory accuracy as the true level of ad familiarity increases. However, unlike $A'$, $H''_C$ confounds bias and accuracy at all levels of true ad memorability. Clearly, this index is not useful.

Panel E presents memory performance for collapsed $A_z$. In contrast to $H''_C$ and the uncorrected hit rate, $A_z$ clearly and consistently separates decision bias from memory sensitivity. Moreover, in contrast to $A'$, $A_z$ maintains its validity even at high levels of true ad familiarity. These results are reinforced by correlation coefficients, displayed in Table 1, computed between true ad familiarity and the five competing measures. Note that $A_z$ correlates best with actual ad familiarity, calculated separately for the four groups and when pooled across those groups. Also note that correlations for the other four measures are lower but ordered in a fashion that would be predicted from data in Figure 1: $d'$ performs better than $A'$ and the uncorrected hit rate, which in turn outperforms $H''_C$.

### *Simulation 2: The Consistency and Efficiency of Collapsed Estimators*

Macmillan and Kaplan (1985) argue that one of the most important precautions in using collapsed measures is to avoid aggregating data across subjects with different response criteria. If considerable variability is present, subjects should be clustered into subgroups with similar bias, and averages computed across those subgroups (Crowder 1982). In single observation designs, it is difficult to determine the underlying bias of individual respondents, thereby precluding aggregating across subgroups. Therefore, the second simulation has two goals. First, though $A_z$ clearly represents the best measure when examined in terms of statistical bias, a remaining concern is the possible impact of criterion variability on the statistical consistency and efficiency of the various indices. Second, because the collapsed measures pool data across a group of respondents, what is the impact on the resulting index if each of those respondents adopts a different response criterion?

*Method*. The general approach of the second simulation is similar to that of the first. Probability density functions corresponding to target and distractor distributions were constructed such that the distractor distribution was assigned a mean of zero, and the target distribution was assigned one of three different mean values, .5, 1.5, and 3.0. Variances for both distributions were set at unity. The goal of the second simulation was to assess the various estimators in the face of individual subject criterion variability. To accomplish this, an overall decisional criterion was set at .3, and five levels of variability around that criterion were modeled ranging from a low of 0 to a high of .8 (0, .2, .4, .6, .8).

Because statistical consistency is defined as a reduction in bias with increasing sample size, six different sample sizes were defined ($n$ = 50, 75, 100, 150, 200, 400). This resulted in a 6 × 5 × 3 factorial design with the six sample sizes, five levels of criterion variance, and three levels of true $d'$. Efficiency was assessed by computing the root mean square error of the various collapsed estimators.

The biggest change from the first simulation is that

Figure 1
SIMULATED RECOGNITION PERFORMANCE

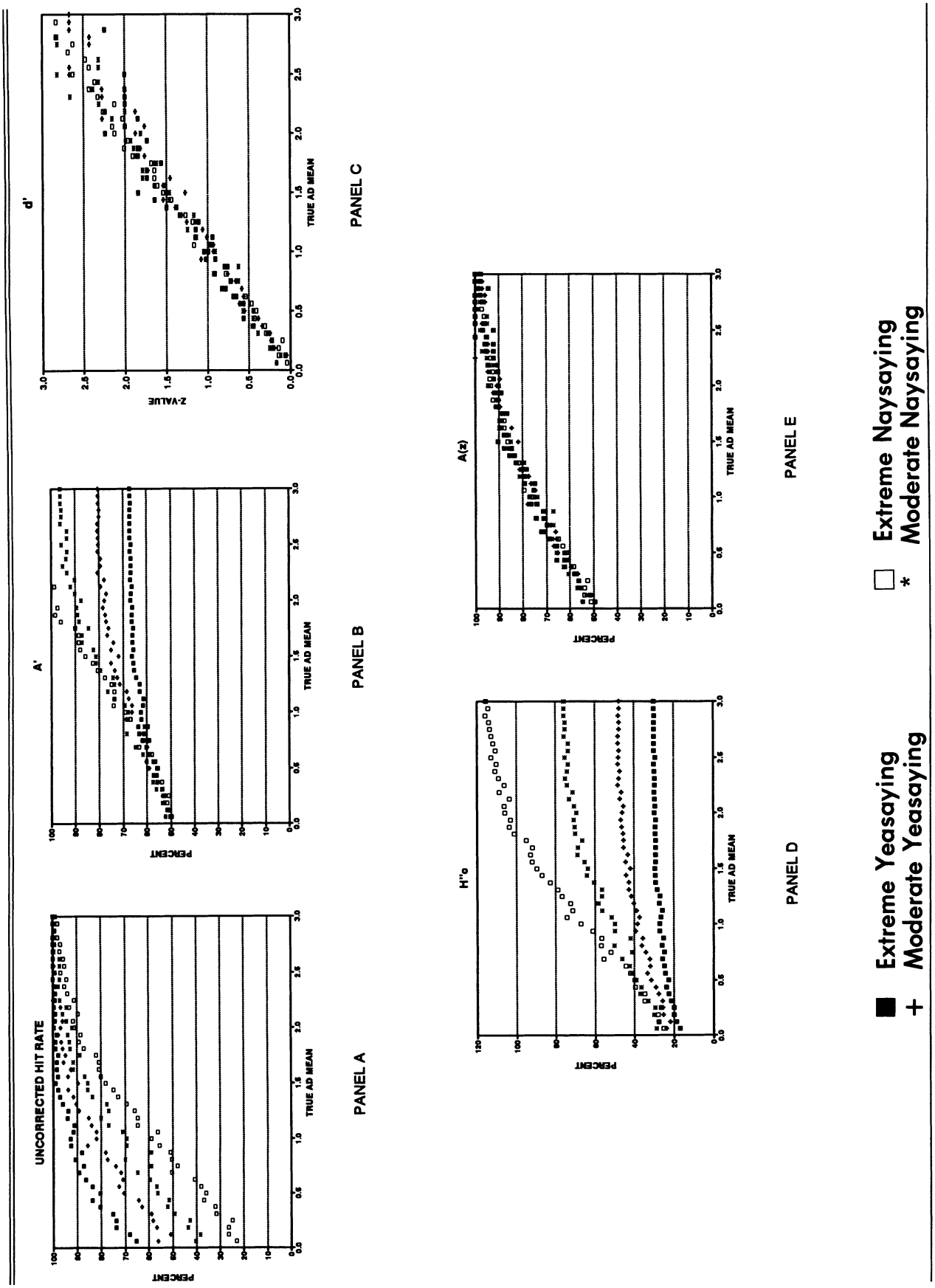

Table 1
CORRELATIONS BETWEEN THE AD MEAN AND FIVE MEASURES OF AD MEMORABILITY (TWO-TAILED PROBABILITY FOR THE CORRELATION APPEARS IN PARENTHESES)

| | Measure | | | | |
|---|---|---|---|---|---|
| *Criterion Group* | *Raw Hit Rate* | $d'$ | $H''_C$ | $A'$ | $A_z$ |
| Pooled | .8101 (.0001) | .8911 (.0001) | .5100 (.0001) | .7667 (.0001) | .9636 (.0001) |
| Extreme Yeasaying | .8629 (.0001) | .8888 (.0001) | .8747 (.0001) | .8935 (.0001) | .9585 (.0001) |
| Moderate Yeasaying | .9240 (.0001) | .8319 (.0001) | .9313 (.0001) | .9503 (.0001) | .9649 (.0001) |
| Moderate Naysaying | .9526 (.0001) | .9880 (.0001) | .9585 (.0001) | .9739 (.0001) | .9697 (.0001) |
| Extreme Naysaying | .9689 (.0001) | .9953 (.0001) | .9710 (.0001) | .9874 (.0001) | .9628 (.0001) |

because of the superior performance of $A_z$ and $A'$, only these were included in the second simulation. Because $A_z$ represents a transformation of $d'$ in the case of one $h$ and $f$ pair, in a sense $d'$ is dominated by $A_z$, and it is not necessary to analyze both measures. In addition, because the value of $H''_C$ depends on the specific distribution of hits and false alarms across individuals, we could not produce a "true" value for the measure without producing "true" values for each subjects' responses. Furthermore, because $H''_C$ behaved so poorly in the first simulation, it seemed pointless to test it for consistency.

*Results*. The results of the simulation are displayed in Table 2, which shows root mean square error for each level of the 6 × 5 × 3 design. Because the upper level of true underlying familiarity was set at 3.0, there is some distortion in the tabled values due to a ceiling effect. For example, when familiarity is set at 3.0, a large number of simulated experiments result in hit rates of 100%, thus precluding the calculation of $A_z$. Though an underlying $d'$ of 3.0 is artificially high, nevertheless, this value was chosen to highlight the impact of high levels of familiarity on the indices. The implication of this for our analysis is that the present simulation presents a worst-performance scenario for $A_z$.

An examination of the tabled values suggests several conclusions. First, the second simulation closely parallels the findings of the first: Error scores for $A_z$ suggest that it can be estimated quite accurately, even at high levels of true memorability. In contrast, the $A'$ error increases as the true level of memorability increases from .5 to 1.0 and is clearly evident as underlying memory reaches 3.0. Second, error for both indices steadily diminishes as the sample size increases from 50 to 400, suggesting that both $A_z$ and $A'$ appear to be relatively consistent. To support this observation, an analysis of variance was conducted on the underlying bias scores (signed deviations) used to compute the root mean square error in Table 2. Statistical consistency should be revealed through a main effect for sample size and a lack of significant interactions between sample size and the other factors in the model. The ANOVA reveals just such a pattern: The analysis of the $A_z$ scores revealed a significant main effect for sample size, $F_{5,\ 2231} = 5.16$, $p < .0001$, and the analysis of the $A'$ estimates showed a similar significant main effect, $F_{5,\ 2610} = 2.74$, $p < .02$. In both cases the average bias in the estimates decreased as the sample increased. Likewise, no significant higher-order interactions involving sample size were found for either $A_z$ or $A'$. Finally, an examination of the values in Table 2 suggests that, though the mean square error is lower for $A'$ than for $A_z$ when the mean is .5, within the parameters of the present simulation, $A_z$ appears to be the more efficient of the two estimates. In summary, the second simulation clearly suggests that $A_z$ appears to be a relatively unbiased, consistent, and efficient estimator. In contrast, though $A'$ performed in a relatively consistent manner, it does appear to be less efficient. Moreover, the second simulation also confirms that $A'$ displays considerable statistical bias at higher levels of true memorability.

The second objective of the simulation was to determine the impact of variable subject criteria on the indices. Subject criterion variance does not appear to have much of an impact on statistical efficiency, as seen in Table 2. The ANOVA conducted on bias scores from individual simulations does reveal that increasing subject criterion variance leads to increasing negative bias for both $A_z$ and $A'$. As already noted, however, because sample size does not interact with criterion variance, this would not seem to present a major problem. To the extent that this is a concern, an obvious solution is to use larger sample sizes.

## Table 2
ROOT MEAN SQUARE ERRORS FOR SIMULATION 2

| Var | Sample size | $A_z$ True ad mean .5 | $A_z$ True ad mean 1.0 | $A_z$ True ad mean 3.0 | $A'$ True ad mean .5 | $A'$ True ad mean 1.0 | $A'$ True ad mean 3.0 |
|---|---|---|---|---|---|---|---|
| | 50 | .065 | .045 | .035 | .051 | .080 | .074 |
| | 75 | .053 | .032 | .025 | .041 | .060 | .071 |
| 0 | 100 | .042 | .029 | .022 | .033 | .051 | .048 |
| | 150 | .035 | .030 | .015 | .026 | .047 | .049 |
| | 200 | .034 | .020 | .008 | .028 | .033 | .043 |
| | 400 | .021 | .015 | .008 | .016 | .025 | .028 |
| | 50 | .081 | .056 | .037 | .058 | .105 | .066 |
| | 75 | .052 | .045 | .036 | .039 | .064 | .063 |
| .2 | 100 | .050 | .036 | .021 | .041 | .065 | .055 |
| | 150 | .037 | .032 | .015 | .029 | .049 | .041 |
| | 200 | .030 | .020 | .012 | .024 | .031 | .041 |
| | 400 | .022 | .015 | .009 | .017 | .025 | .032 |
| | 50 | .066 | .046 | .040 | .049 | .067 | .096 |
| | 75 | .061 | .037 | .036 | .047 | .054 | .067 |
| .4 | 100 | .040 | .035 | .021 | .030 | .047 | .064 |
| | 150 | .035 | .035 | .017 | .025 | .050 | .043 |
| | 200 | .036 | .024 | .012 | .026 | .032 | .030 |
| | 400 | .018 | .019 | .012 | .014 | .027 | .031 |
| | 50 | .077 | .060 | .038 | .053 | .076 | .089 |
| | 75 | .041 | .055 | .029 | .029 | .070 | .070 |
| .6 | 100 | .036 | .057 | .034 | .025 | .077 | .055 |
| | 150 | .046 | .048 | .025 | .032 | .060 | .048 |
| | 200 | .036 | .046 | .026 | .025 | .060 | .049 |
| | 400 | .025 | .041 | .020 | .018 | .051 | .036 |
| | 50 | .059 | .081 | .053 | .044 | .096 | .080 |
| | 75 | .049 | .078 | .043 | .034 | .090 | .071 |
| .8 | 100 | .053 | .071 | .038 | .036 | .086 | .048 |
| | 150 | .039 | .074 | .033 | .028 | .089 | .057 |
| | 200 | .036 | .064 | .040 | .026 | .078 | .058 |
| | 400 | .034 | .063 | .035 | .024 | .078 | .055 |

### *SUMMARY AND RECOMMENDATIONS*

We tested the statistical bias, consistency, and efficiency of four different SDT sensitivity measures: a corrected-hit probability measure suggested by Singh and Churchill (1986), the traditional $d'$ statistic, and two nonparametric measures collected from a collapsed-data procedure suggested by Macmillan and Kaplan (1985). Using simulated data, we clearly show the superiority of $A_z$ and $d'$ to the other measures. $A_z$ is somewhat nonlinear with true ad memorability, whereas $d'$ is somewhat more variable as memorability levels increase toward perfect performance. $A'$ also showed a positive relationship between actual and estimated memorability but, unlike $A_z$, confounded decisional bias and sensitivity at higher levels of true ad memorability in a manner systematically linked to yeasaying/naysaying. Finally, $H''_C$ produced the least valid estimates of ad sensitivity. The measure confounded decisional bias and sensitivity across all levels of true ad memorability. Furthermore, the formal evaluation of $H''_C$ presented in the Appendix suggests that it could violate fundamental assumptions of SDT models. A second simulation tested the consistency and efficiency of the $A_z$ and $A'$ measures by manipulating the variability of the decisional criteria adopted by individual respondents within the group. The results showed that the $A_z$ measure is remarkably consistent and, in the comparison with $A'$, relatively efficient. Overall, our results reinforce similar evidence from Macmillan and Kaplan (1985) that collapsed procedures produced relatively unbiased and efficient estimators.

#### *Recommendations*

Though further testing of the collapsed $A_z$ and $A'$ measures is warranted, we can draw some relatively clear recommendations for their use in ad testing. In the majority of cases, the best approach to using SDT for ad recognition testing is to employ confidence-rating pro-

cedures (as discussed in footnote 2) and compute $A_z$ on the resulting collapsed data. This approach is relatively straightforward and should produce few burdens on typical ad testing methodologies currently in use. A FORTRAN-based program is available (Dorfman and Alf 1969; see Swets and Pickett 1982 for a program listing) that can calculate easily the $A_z$ measure as well as a variety of additional statistics from confidence ratings and allow the researcher to examine carefully the collapsed ROC for a particular target ad. Though computationally more difficult, multiple probabilities allow the researcher to assess differential signal and noise variances. When such differential variance exists, single-point summaries can be biased.

Though the $A_z$ measure represents the most appropriate statistic for SDT analyses of ad recognition data, researchers may wish to rely on the computationally simpler $A'$. As was evident from the first simulation, however, one should use caution in employing the measure when the true level of ad familiarity is suspected to be quite high ($d'$s greater than 1.0). This corresponds to an $A'$ score in the range of 60% to 70%.

At this point it is important to note that the previously tested indices all have been computed from data collected from a yes/no paradigm. We have limited our considerations to this paradigm because it enables easy collection of data and it has a high level of ecological validity (e.g., consumers typically react to one ad at a time). In addition, yes/no paradigms enable the researcher the ability to directly measure response bias and to test the hypothesis that target and distractor ads have similar error variance. Though $A'$ and the yes/no paradigm are computationally simple, if response bias is not of interest in and of itself, then ad researchers may want to consider the use of the two-alternative, forced-choice paradigm (see Tashchian, White, and Pak 1988 for a discussion). In a 2AFC paradigm, the percent-correct index of recognition performance is equivalent to $A_z$.

Some mention is necessary regarding the selection of the distractor and target ad similarity. One can imagine easily that the selection of distractor ads highly similar to the target might produce apparently low levels of recognition because the two are so similar. Likewise, one could produce an artificially high measure of recognition by ensuring that the distractors are grossly dissimilar from the target. We recommend that the distractor ads be chosen with an eye to realism. In particular, the distractors should be representative of the actual ads exposed to the target market in the target medium. In this way the separation of the target and distractor distributions will correlate most closely with the familiarity of the target ad against the kind of noisy background actually confronting the consumer as he or she queries memory.

Finally, it is important to note certain limitations with the collapsed index approach. Though this methodology should be the preferred procedure for researchers interested in estimating ad familiarity in a laboratory session, the assumptions of this procedure should be explored prior to any field applications. SDT is designed to be employed within strict bounds of experimental control in which the actual exposure history of each respondent with each target stimulus is known in advance. Without this information, it is difficult to conclusively determine whether a real ad represents signal or noise.

## APPENDIX
## *DERIVATION OF THE ROC FOR THE BIAS-ADJUSTMENT MEASURE*

A common strategy when investigating the validity of an index of memorability is to derive its ROC (e.g., Grier 1971; Swets 1986). By utilizing the definition of the index, one can solve for the hit rate of a specific real ad as a function of its false alarm rate.

To begin, we define $y_{ij}$ as the dummy variable that is set to 1 if subject $i$ ($i = 1, 2, \ldots N$) claims to remember reading ad $j$ ($j = 1, 2, \ldots, n$); 0 otherwise. Also, we define $x_{ik}$ as the dummy variable that is equal to 1 when subject $i$ claims to remember distractor ad $k$ with $k$ also varying from 1 to $n$. Note that our results do not require an equal number of real and distractor ads.

Using our notation the hit rate for ad $j$ is simply

$$h_j = \frac{1}{N}\sum_{i}^{N} y_{ij} .$$

The pooled false alarm rate is

$$f = \frac{1}{Nn}\sum_{1}^{N}\sum_{k}^{n} x_{ik}$$

or equivalently

$$f = \frac{1}{N}\sum_{i}^{N} f_i ,$$

where $f_{i.}$ is defined as the false alarm rate for subject $i$, that is,

$$f_{i.} = \frac{1}{n}\sum_{k}^{n} x_{ik} .$$

The hit rate for subject $i$ is

$$h_{i.} = \frac{1}{n}\sum_{m}^{n} y_{im} .$$

Assume $h_{i.} + f_{i.} \geq 1$ for all $i$, which implies that subjects are yeasaying. In that case we can write the Singh and Churchill (1987) measure for ad $j$ as

$$H_c'' = h_{.j} - \frac{1}{N}\sum_{i}^{N}\left[\frac{h_{i.}(1 - h_{i.})}{f_{i.}(1 - f_{i.})} - 1\right] y_{ij} .$$

Note that the term in the brackets is the measure of bias from Hodos (1970) assuming yeasaying. Some algebra leads to

$$h_{.j} = \frac{1}{2}H_c'' + \frac{1}{2N}\sum_{i}^{N}\frac{h_{i.}(1 - h_{i.})\, y_{ij}}{f_{i.}(1 - f_{i.})} .$$

To express $h_{.j}$ as a function of f, we could redefine the $f_{i.}$ in the preceding expression as follows:

$$f_{i\cdot} = Nf - \sum_{i' \neq i} f_{i'} \ .$$

If this expression is substituted for $f_{i\cdot}$ into the equation for hit rate, however, a value of f would not uniquely determine $h_{\cdot j}$. Instead, there is a whole family of ROC curves depending on the specific distribution of false alarms, and hits, across the individuals in the study.

Now consider the case in which subjects are naysaying. In that case the bias adjustment measure is

$$H_c'' = h_{\cdot j} - \frac{1}{N}\sum_{i}^{N}\left[1 - \frac{f_{i\cdot}(1 - f_{i\cdot})}{h_{i\cdot}(1 - h_{i\cdot})}\right] y_{ij}$$

$$= \frac{1}{N}\sum_{i}^{N}\frac{f_{i\cdot}\,(1 - f_{i\cdot})y_{ij}}{h_{i\cdot}\,(1 - -h_{i\cdot})} \ .$$

As we now see, different ROCs are implied for naysaying and yeasaying samples, which violates an important assumption of the theory of signal detection; namely that an ROC represents constant memorability with response bias varying within one ROC curve. Therefore, the bias adjustment measure is inconsistent with the theory of signal detection.

## REFERENCES


Appel, Valentine and Milton L. Blum (1961), "Ad Recognition and Respondent Set," *Journal of Advertising Research*, 1 (4), 13–21.

Bagozzi, Richard P. and A. J. Silk (1983), "Recall, Recognition, and the Measurement of Memory for Print Advertisements," *Marketing Science*, 2 (2), 95–134.

Banks, W. P. (1970), "Signal Detection Theory and Human Memory," *Psychological Bulletin*, 74 (August), 81–89.

Crowder, Robert G. (1982), "A Common Basis for Auditory Sensory Storage in Perception and Immediate Memory, *Perception and Psychophysics*, 31, 477–83.

Dorfman, Donald D., and Edward Alf, Jr. (1969), "Maximum-Likelihood Estimation of Parameters of Signal Detection Theory and Determination of Confidence Intervals-Rating Method Data," *Journal of Mathematical Psychology*, 6 (October), 487–96.

Fisk, A. D. and Schneider, W. (1984), "Memory as a function of attention, level of processing, and automatization," *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 10 (January), 181–97.

Green, David M. and John A. Swets (1966), *Signal Detection Theory and Psychophysics*. New York: John Wiley & Sons, Inc.

Grier, J. Brown (1971), "Nonparametric Indexes for Sensitivity and Bias: Computing Formulas," *Psychological Bulletin*, 75 (June) 424–9.

Hodos, W. (1970), "Nonparametric Index of Response Bias for Use in Detection and Recognition Experiments," *Psychological Bulletin*, 74 (November), 351–4.

Leigh, James H., and Anil Menon (1986), "A Comparison of Alternative Recognition Measures of Advertising Effectiveness," *Journal of Advertising*, 15 (3), 4–20.

Lucas, D. B. (1942), "A Controlled Recognition Technique for Measuring Magazine Advertising Audiences," *Journal of Marketing*, 6 (4), 133–6.

Macmillan, Neil A. and Howard L. Kaplan (1985), "Detection Theory Analysis of Group Data: Estimating Sensitivity From Average Hit and False-Alarm Rates," *Psychological Bulletin*, 98, 185–99.

McNicol, D. (1972), *A Primer of Signal Detection Theory*. London: Allen & Unwin.

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 (November), 327–36.

——— and ——— (1987), "Response-Bias-Free Recognition Tests to Measure Advertising Effects," *Journal of Advertising Research* (June/July), 23–36.

——— and M. L. Rothschild (1983), "Recognition as a Measure of Learning from Television Commercials," *Journal of Marketing Research*, 20 (August), 235–48.

Swets, John A. (1986a), "Indices of Discrimination or Diagnostic Accuracy: Their ROCs and Implied Models," *Psychological Bulletin*, 99 (January), 100–17.

——— (1986b), "Form of Empirical ROCs in Discrimination and Diagnostic Tasks: Implications for Theory and Measurement of Performance," *Psychological Bulletin*, 99 (February), 181–98.

——— and Ronald M. Pickett (1982), *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press.

Tashchian, Armen, J. Dennis White, and Sukgoo Pak (1988), "Signal Detection Analysis and Advertising Recognition: An Introduction to Measurement and Interpretation Issues," *Journal of Marketing Research*, 24 (Nov), 397–404.

Wells, Williams D. (1961), "The Influence of Yea-Saying Response Style," *Journal of Advertising Research*, 1 (June), 1–12.