SURENDRA N. SINGH and GILBERT A. CHURCHILL, JR.*

Recognition tests are a very popular means of assessing the memory effectiveness of advertisements. Unfortunately the recognition scores obtained by current methods reflect both the memory for an advertisement and the response biases of the respondents. The authors introduce the theory of signal detection (TSD) which can be used to secure independent estimates of memory and response bias in recognition tests. They discuss how TSD can be used to improve ad recognition testing.

Using the Theory of Signal Detection to Improve Ad Recognition Testing

Of the several methods for measuring the impact of printed advertisements, one of the most widely used is the recognition method (Singh and Rothschild 1983). "The critical thing that defines a recognition task is that the person is given a *copy* of the information he or she needs to find in memory" (Glass, Holyoak, and Santa 1979, p. 59). In a typical recognition test, such as performed by Starch/INRA/Hooper, subjects are shown a series of advertisements, one at a time. As each advertisement appears, the subjects are to respond "yes" if they think they have seen the ad earlier and "no" if not. In spite of their popularity, these recognition tests are shrouded in controversy. Appel and Blum (1961) and Marder and David (1961) pointed out long ago, for example, that a large percentage of respondents will claim recognition of bogus ads (ads respondents could not have seen before) contained in magazines when real ads are also being tested. In some studies, the claimed level of recognition for bogus ads has been almost as high as that for real ads (Simmons 1961).

The tendency to "recognize ads," irrespective of prior exposure to them, may be due to "acquiescence response set" bias, that is, the general tendency to favor affirmative responses over negative responses apart from the content of the items at issue. In yes/no or true/false tests, for example, more individuals give an excess number of "yes" or "true" responses (Cronbach 1950; Guilford 1967). Acquiescence set bias has been called "noting set" (Appel and Blum 1961) or "yea-saying" (Wells 1961) bias in the advertising context.

Noting set bias is only one variable contributing to excessively high ad noting scores. Among the other factors that affect ad recognition scores are guessing when uncertain, eagerness to please the interviewer, hesitation to appear ignorant, and the tendency to deny socially undesirable traits and to admit to socially desirable ones (Clancy, Ostlund, and Wyner 1979; Lucas and Britt 1963).

Several methods have been suggested over the years for making recognition tests more valid. Most of these methods rely on the inclusion of "false" (bogus) ads along with true "stimulus" ads in the test. The methods differ primarily in terms of how the distractor ads are used. One variation exposes a control group of subjects to a portfolio of distractor ads to develop a benchmark of the degree of false claiming, which is then used to adjust the claimed recognition scores of subjects exposed to the stimulus ads (Appel and Blum 1961; Lucas 1942; Simmons 1961). Another variation involves informing respondents about the presence of bogus ads to make them aware that they cannot indiscriminately claim recognition of items (Clancy, Ostlund, and Wyner 1979; Neu 1961). A third variation directs respondents to choose the stimulus ad from among one or more distractor ads (Moran 1951a,b; Singh and Rothschild 1983). Unfortunately, all of these methods have been found to have major weaknesses—increased cost, adjusted recognition

*Surendra N. Singh is Assistant Professor of Business, University of Kansas. Gilbert A. Churchill, Jr., is Donald C. Slichter Professor in Business Research, University of Wisconsin–Madison.

This project was supported in part by the University of Kansas General Research Fund allocation #3037-20-0038. Support also was given by the University of Kansas School of Business Research Fund provided by the Volume Shoe Corporation. The ideas and opinions expressed herein are solely those of the authors. The authors thank the editor and the five anonymous *JMR* reviewers for their many helpful comments.

Journal of Marketing Research
Vol. XXIII (November 1986), 327–36

scores that are negative (a logical impossibility), and the fact that the choice of distractor ads can affect the obtained scores. In effect, none of the methods is able to produce a recognition measure that could be considered a true indicator of recognition memory.

The purpose of our article is to present the basics of signal detection theory, which offers promise in providing a measure of recognition memory uncontaminated by the response tendencies of subjects. We begin with a brief review of the theory of signal detection (TSD). Next we discuss the results from two experiments that examine some propositions from TSD in an advertising context. Finally, we describe the implications and limitations of the theory of signal detection when used to assess the effectiveness of print ads.

THEORY OF SIGNAL DETECTION

TSD originated in World War II as psychologists were trying to make ground observers of enemy planes more accurate. It later was used in electrical engineering to aid in the design of sensing devices (Peterson, Birdsall, and Fox 1954; Van Meter and Middleton 1954), as well as in many areas of psychology because it can be applied to any situation where sensory input is ambiguous. (For a general overview of the many types of situations to which TSD has been applied, see Green and Swets 1966).

The key notions in TSD can be understood from its application to recognition testing of memory. In a typical recognition memory test, the subject is presented with a list of items (e.g., words) one at a time, some of which he or she has been exposed to in an earlier session and others of which are distractors. As each item is presented, the subject is to respond "yes" if he or she thinks that the item was on the original list and "no" if not. Subjects are told beforehand the *proportions* of items that were on the original list and of items that are distractors. Items to which subjects have been exposed previously should be familiar to them; in signal detection language, old or familiar items are called "signals" or "stimulus items" and new or distractor items are called "noise" (Banks 1970). Subjects can be paid for every correct response and can be penalized for every incorrect response, typically by withholding the reward. Usually nonmonetary rewards (e.g., eagerness to please the interviewer, hesitation to appear ignorant, and so on) also are operating which affect the answers given by the subject.

There are four possible outcomes to every trial—the subject may say either that the word was familiar or that it was new and the trial may have been either signal or noise. A *hit* response is one in which the subject says "yes" to the presence of a signal and signal was actually present; a *miss* occurs when the subject says "no," but the signal was present; a *false alarm* occurs when the subject reports the presence of a signal but in reality the trial contained noise alone; and a *correct rejection* occurs when the subject says no signal was present and the trial actually did not have a signal. Notice that the sum of the probabilities of hit and miss must equal 1.00; similarly, the sum of the probabilities of false alarm and correct rejection must add to 1.00. Alternatively, the probability of total yes and no responses given the signal was or was not present must sum to 1.00. Because of this complementary relationship between the responses, only two responses customarily are used to summarize a subject's performance—the hit and false alarm ratios.

The performance of a subject in the recognition test depends mainly on two factors, the subject's ability to perform the task and the subject's motivational state and response tendencies (Pastore and Scheirer 1974). The experimenter can affect the subject's response tendencies and motivations by changing payoffs and/or by changing prior odds. For example, in a word recognition test, if the subject is aware that there is no penalty for incorrect answers, he or she would probably have a greater motivation for guessing than if wrong answers were scored negatively. However, the subject's discrimination ability should remain unaffected by changes in motivational factors. Unfortunately, these two aspects—the sensory or discrimination capabilities of the subject and his or her decision-making style (e.g., the effect of his or her values, motivations, knowledge of prior odds, and so on)—are completely confounded in the responses secured from the subject.

> The basic aim and unique contribution of the TSD is the separation of the sensory capabilities of the subject from the individual's decision-making aspects and the precise estimation of each (Coombs, Dawes, and Tversky 1970, p. 166).

Assumptions of TSD in the Study of Memory

The first assumption of TSD is that any information an individual possesses has a certain strength in long-term memory. The strength of the item can be taken as the strength of a memory trace for it. Alternatively, it can refer to the degree of familiarity; the more familiar an item is, the greater would be the memory strength for it and vice versa. A second convenient assumption for the time being is that measurements of the strength of items, both old and new, are normally distributed and have equal variances. This means in essence that there are two normal distributions for subjects to consider, one representing the list of familiar items and one representing the list of distractor items. The assumptions of normality and equal variance were made in the original development of the theory and are useful for illustrating the essential notions.[1] They are no longer necessary, however, because the *key statistics* that come out of the normal theory approach have parallels requiring *no assumptions* about the underlying distributions. Consequently, we use the normal distribution and equal vari-

[1]For a theoretical argument as to why the distributions should be normal, see Egan and Clarke (1966).

ance assumptions to highlight the basic arguments and key interpretive quantities of TSD, though we use their distribution-free counterparts in discussing the results of our experiments.

Finally, TSD assumes that an individual's exposure to an item increases its strength in long-term memory. In other words, both the stimuli and distractor items have certain strength value to begin with, but the strength value is changed with exposure to the item during the experiment. Consider the full set of items making up a test. Some of them might have been very familiar to the subject previously, some might have been very unfamiliar, and still others might have been moderately familiar. Because the subject is exposed to the stimulus items (but not the distractor items) before the test, they increase in strength value in comparison with the distractor items, which remain at their initial strength. In effect, the distribution of old items on the familiarity continuum is moved to the right by a fixed amount (for the rationale supporting these assumptions, see Klatzky 1980).

When faced with an item on a recognition task, the subject must decide whether the item is old (stimulus) or new (distractor), that is, did it come from the distribution of old items $f_o(x)$ or from the distribution of new items $f_n(x)$. Figure 1,A depicts the situation.

The familiarity continuum is plotted on the abscissa in Figure 1; $f_o(x)$ is to the right of $f_n(x)$ because most old items would have greater familiarity value (or memory strength) than new items. There is some overlap between the distributions, however, because though the mean strength of old items would be greater than that of new items, some new items may have higher memory strength than some old items.[2] The distance between the means of the two distributions is a measure of how far apart the two are on the familiarity continuum. Each familiarity value x has two probability densities attached to it—one from the new and one from the old distribution. Hence, each x value on the familiarity continuum may be associated with a particular likelihood ratio, $1(x)$, defined as

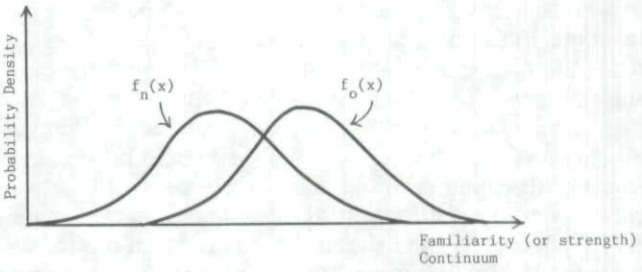
$$1(x) = f_o(x)/f_n(x) = \frac{p(x/o)}{p(x/n)}.$$

The likelihood ratio thus reflects the likelihood that a particular item belongs to the class of old items relative to the likelihood that it belongs to the class of new items (Swets, Tanner, and Birdsall 1964).

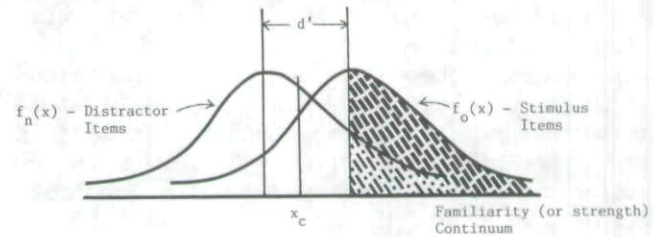
> Notice that while the numerator and denominator of the likelihood ratio are the probabilities of an observation under two different hypotheses, the likelihood ratio is a number not a probability. The number is a function of whatever variables were involved in the observation, but $1(x)$ is itself only a *single* variable. We may say, then, that the *decision axis* becomes the likelihood ratio and the axis is continuous and unidimensional (Corso 1967).

Figure 1
SOME KEY NOTIONS ABOUT THE DISTRIBUTIONS OF OLD AND NEW ITEMS

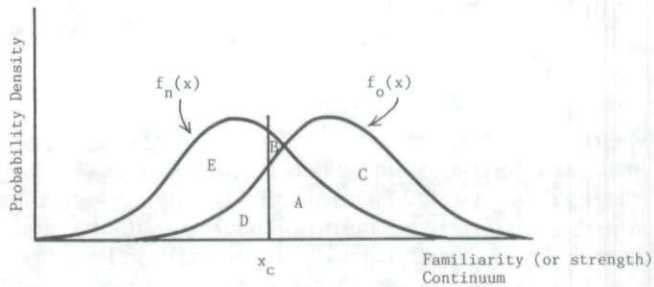
A. Probability Density Functions of Old and New Items



B. Impact of Cutoff Value x_c (Critical Value of x Corresponding to $1(x) = \beta$) Given a Certain Set of Circumstances



C. Visual Representation of D'



According to TSD, in other words, a subject does not base his or her decision on the value of the raw sensory input x, but rather on a transformation of it to a new decision axis, the likelihood ratio. When the subject is presented with an item to which he or she must respond "old" or "new," the individual acts as though he or she were computing the likelihood ratio associated with some

[2] In an advertising recognition test the question being asked is, "Did you ever encounter this ad in this particular issue of this publication (magazine/newspaper, etc.)?" This is essentially a test of *episodic memory instead of semantic memory*, for which the appropriate question would be, "Have you ever encountered this ad before?" (Wallace 1980). Thus it is possible that a subject may be more familiar with certain distractor ads because of prior exposure to them in publications other than the one being evaluated.

fixed *criterion* value of $1(x)$ called β. If $1(x) \geq \beta$, the subject responds "old" or "yes, the item is familiar"; otherwise "new," or "no, I have not seen the item before." β thus represents a threshold for saying "yes."

A similar argument can be developed for ad recognition experiments. Consider, for example, the question of whether a subject has been exposed to a particular ad in a specific issue of a magazine. When faced with a recognition task, the subject must decide whether he or she has seen the item or advertisement before in the magazine issue in question, that is, whether the item is an old or stimulus item. Now the item could have come from the distribution of old items, $f_o(x)$, or it could have come from the distribution of new items, $f_n(x)$. Figure 1,A again captures the situation; $f_o(x)$ is again plotted to the right of $f_n(x)$ because most ads in the test issue would have greater familiarity value (or memory strength) than items not in the issue if the subject was indeed exposed to the issue, though there could be some overlap between the distributions. The distance between the means of the two distributions is a measure of how far apart the two are on the familiarity continuum. The likelihood ratio expresses the likelihood that the particular item belongs to the class of old items (actual ads) versus the likelihood that it belong to the class of new items (distractor ads), and if $1(x) \geq \beta$, the subject responds that he or she did see the ad in the magazine issue in question. As a threshold for saying "yes," β can be shown to be a function of the individual's response biases, attitudes, and motives along with the prior probabilities of the occurrence of the stimulus item in a given test. More particularly, it can be shown that (see Coombs, Dawes, and Tversky 1970, p. 170–2)

$$\beta = k \frac{p(n)}{p(o)}$$

where k is a constant of a proportionality and $p(n)$ and $p(o)$ are prior probabilities that an item is a new (distractor) or old (stimulus) item, respectively. To see the impact of a change in the prior odds of stimulus and distractor items on β, consider Figure 1,B and assume $k = 1$.[3] Assume further that the subject is using x_c (the critical value of x corresponding to $1(x) = \beta$) as a cutoff, that is, the subject is saying "yes, I have seen the ad before" for all those items for which $x > x_c$. The area denoted A + C under the old distribution represents the proportion of hits and the area A + B under the new distribution represents the proportion of false alarms; area D represents the proportion of misses and area E represents the proportion of correct rejections. When there are equal numbers of stimulus and distractor items, $\beta = 1$ and the critical value of x (i.e., x_c) is the point where $f_o(x) = f_n(x)$, that is, where the two distributions intersect. Suppose, however, there are nine distractor items and one stimulus item; the prior odds for an old item would be 1 in 9, β would be 9, and $1(x)$ would have to be greater than or equal to 9 before the subject would say "yes, the item is old." Similarly, if the odds are 60:40 in favor of old items, β would be .67 and $1(x) \geq .67$ before the subject would say "yes, the item is old." In sum, when the odds are more in favor of old items, a *lower* β is needed for a subject to say "yes" than when the odds are against them. Conversely, if the prior odds are unfavorable for the occurrence of a stimulus ad, the evidence must be more substantial for a subject to conclude the stimulus ad is present.

In TSD terms, the β parameter is said to reflect the subject's response tendencies and motivational states, for example, whether the subject is aggressive for some reason in saying "yes" or has yea-saying tendencies, or is cautious or has nay-saying tendencies. It does not measure the true discrimination ability or sensory capabilities of the subject. Rather, the individual's actual recognition memory is unaffected by his or her response tendencies, but does change as the subject's memory capabilities change. According to TSD the distance between the means of the two distributions $f_o(x)$ and $f_n(x)$ provides a measure of *sensitivity* of the sensory system. This distance, designated as d' (d-prime), is independent of the decision criterion used by the subject and changes only when there is a change in the subject's true sensitivity with respect to his or her discrimination ability between old and new items. Because d' is the distance between the means of the two distributions $f_o(x)$ and $f_n(x)$, which are assumed to be normal with equal variance (see Figure 1,C), d' can be calculated by the expression

$$d' = z_{FAR} - z_{HR}$$

where:

z_{FAR}, the normalized deviate for the false alarm rate, equals the number of standardized score units the mean of the distractor item distribution is away from x_c, and

z_{HR}, the normalized deviate for the hit rate, equals the number of standardized score units the mean of the old item distribution is away from x_c.

A $d' = 0$ implies no difference between the means of the distributions of the old and new items; in essence the distributions lie on top of each other, indicating the subject cannot discriminate at all between the two types of stimuli. A $d' > 0$ implies the individual is basically able

[3]The assumption that $k = 1$ implies that the payoffs for hits and correct rejections are equal and the penalties for false alarms and incorrect rejections are also equal. That is, correct decisions of each type are rewarded equally and errors, regardless of type, are penalized equally. In such a case, the individual would be motivated to be as accurate as possible. The argument applies to both monetary, if any, and psychic rewards. Thus if the respondent for some psychic reason valued hits more than correct rejections, he or she would be more inclined to say "yes"; k would no longer be 1 and the threshold for saying "yes" would shift as well.

to discriminate between the old and new stimuli. The greater the d', the better the discrimination. A $d' < 0$ implies either (1) measurement error or (2) the subject is performing the discrimination task and then giving a contrary response, that is, saying "no" when he or she should have said "yes" (and vice versa) on the basis of discrimination ability. In other words, the subject is able to perform the discrimination task, but is malingering (Pastore and Scheirer 1974).

Estimating β and d'

Let us suppose a subject is shown 100 ads contained in a folder, one at a time. One half of the ads are stimulus ads in that they appeared in the magazine issue in question and the other half are distractors. The subject is made aware of the prior odds by being told how many ads there are of each type. Suppose further that the subject's hit rate and false alarm rate in the recognition task are 80% and 30%, respectively.

Consider what the subject's hit rate of .80 and false alarm rate of .30 imply. The hit rate means that 80% of the area under the old distribution is to the right of x_c, which is the critical value of the observation corresponding to β. Similarly, the false alarm rate means that 30% of the new distribution is to the right of x_c (see Figure 1,C).

The value of β can be estimated by looking at the ratio of the ordinates at x_c, that is, β = ordinate (HR)/ordinate (FAR). Because the distributions are assumed to be normal, the ordinates at x_c can be obtained from normal probability tables and in this particular instance are ordinate (HR) = .27996 and ordinate (FAR) = .34769. Hence,

$$\beta = \frac{\text{Ord (HR)}}{\text{Ord (FAR)}} = \frac{.27996}{.34796} = .8052.$$

Consider next the estimation of the discriminability index d'. It can be estimated by calculating the difference $z_{FAR} - z_{HR}$, or the distance between the criterion point x_c and the means of the new and old distributions, respectively.[4]

Because 30% of the area of the new distribution in the example is to the right of x_c, 20% of the area must be to the left, between it and the mean of the new distribution. Hence the distance between the mean of the new distribution and x_c corresponds to a z score of .524 (i.e., $z_{FAR} = .524$). Similarly, because 80% of the area under the old distribution is to the right of x_c, 30% of it must be between x_c and the mean of the old distribution. This corresponds to a z score of −.842 (i.e., $z_{HR} = -.842$). Thus, the distance between the two means $d' = z_{FAR} - z_{HR} = .524 + .842 = 1.366$. In other words, the capability of the subject to discriminate between the two classes of events is inversely proportional to the total area common to the two conditional probability density functions.

In sum, TSD provides two parameters, d' and β; d' is a measure of the true memory capability of the respondent for a set of items whereas β is a measure of the subject's response tendencies. Further, the two parameters can vary independently such as when there is no real change in the subject's memory capability but there is a change in motivation level (Banks 1970).[5]

> When β is very low, an item needs little strength for him or her to say "old." Consequently, the response will be "old" very often, with the subject being correct on most of the items that are actually old but committing many errors on new items. In short, there will be a high hit rate, but also a high false-alarm rate. If β is high, the situation is reversed. The subject is very cautious, and seldom says "old" unless he or she is quite sure—which will be only for items with high familiarity. There will be a relatively low hit rate, for the subject will often say "new" to old items, simply because of caution. On the other hand, the subject will also have a low false-alarm rate, because the response "old" will not often be given to new items. Thus we see that if d' remains constant, shifts in β will cause both the hit and false-alarm rates to change, and in the same direction. As β goes up, both hit and false-alarm rate go down (Klatzky 1980, p. 249–51).

Similarly, whenever there is a real increase in the memory for a given set of items (as would be expected if the items were presented over and over) and there is no change in the motivational state of the individual from one recognition test to the next, there would be an increase in the hit rate but not the false alarm rate. The reason is that the respondent's ability to discriminate between the old and new items would truly increase.

Suppose the distributions neither are normal nor have equal variances.[6] Then d' and β are no longer the appropriate measures of the subject's ability to discriminate between the classes of events and the subject's response biases, respectively. The statistics that are appropriate depend on which assumptions do hold. When the distributions are normal but do not have equal vari-

[4]Both d' and β values also can be obtained by using certain standard published tables (see, e.g., Elliott 1964).

[5]If the equal-variance Gaussian model applies, the two parameters are necessarily independent. The normal and equal variance assumptions can be tested by using rating scale responses of the type used in the empirical examples that follow. When the assumptions do not hold, other statistics with meanings similar to d' and β are calculated from a subject's responses. For the description of the procedures for testing the assumptions and the statistics that should be used when either or both assumptions do not hold, see Pastore and Scheirer (1974). For a graphic demonstration of the independence of d' and β, see Klatzky (1980, p. 249–51).

[6]Pastore and Scheirer (1974) provide a useful overview of the statistics that should be used to estimate the discrimination abilities and response biases of subjects under various conditions.

ances, a parallel statistic d'_s is used in place of d' while β remains the same. When one does not want to make any assumptions about the shape of the distributions, one should use nonparametric measures of sensitivity and response bias (Green and Swets 1966). More specifically, the nonparametric measure A' can be used to assess the subject's discrimination abilities and B'_H to assess his or her response tendencies. These parameters, too, can be estimated solely from the subject's hit and false alarm rates in a recognition task and their interpretation is straightforward. Simple computation formulas for both of these measures are given by Grier (1971).[7] More specifically,

$$A' = 1/2 + (y - x)(1 + y - x)/4y(1 - x),$$

$$B'_H = 1 - x(1 - x)/y(1 - y) \text{ for nay-sayers, and}$$

$$B'_H = y(1 - y)/x(1 - x) - 1 \text{ for yea-sayers}$$

where y = hit rate and x = false alarm rate. A' can range from .5 to 1.0 where .5 indicates zero recognition memory or chance performance and 1.0 perfect recognition memory. B'_H scores can range from -100% to $+100\%$ where -100% represents maximum yea-saying and $+100\%$ maximum nay-saying. $B'_H = 0$ indicates unbiased response. Given our very limited current knowledge about the shape of the two response distributions, the nonparametric measures A' and B'_H seem particularly valuable for ad recognition experiments.

EXPERIMENT 1

To examine the potential usefulness of TSD for ad recognition testing, we designed two experiments to test several propositions. The propositions included both attributes of the measures themselves and some expected theoretical relationships.

Research Design

Experiment 1 was conducted among MBA students at a large midwestern university. Students in the experiment were given a portfolio of 48 ads and were asked to evaluate each ad on a 7-point scale using descriptors that ranged from "extremely rational" to "extremely emotional." The disguise was used to help produce recognition memory decline over a relatively short period of time because prior research indicates a relatively slow decay in recognition memory.

Subjects were given another portfolio of ads to evaluate three weeks later. Each of the new portfolios contained the original 48 stimulus ads, but also 48 distractor ads representing similar products, and was divided into two parts. The first half had 24 stimulus and 24 distractor ad pairs mixed at random. The second half also contained 24 stimulus and 24 distractor ad pairs, but ordered in the same sequence as the pairs in the first half. Also, the ads in the first and second halves of the portfolio were matched in terms of product category; thus, if the randomly determined ad sequence had a cigarette ad first, the first ad pair in the second half also involved a cigarette. Subjects were not made aware of this similarity between the two halves of the portfolio. To them, it was one portfolio with 48 stimulus and 48 distractor ads.

Subjects were asked to identify which ad in each pair they had seen three weeks earlier when evaluating ads as to their "emotionality-rationality." Subjects also were asked to indicate the confidence they had in each recognition judgment, using a 5-point confidence rating scale with end anchors "absolutely confident" and "absolutely not confident." Subjects also provided an estimate of the number of magazines they read on a regular basis.

Thirty-two subjects completed both evaluations. The responses of three subjects had to be eliminated because those three had guessed the purpose of the experiment when first asked to evaluate the emotionality-rationality of the stimulus ads. The following analysis is based on the responses from 29 subjects.

Results

Independence of the sensitivity and bias indices is important if one is to differentiate between the subject's sensory capabilities and response tendencies. To examine this issue, A' and B'_H scores were computed for each subject by the formulas presented before and the correlation in indices was computed across subjects. The correlation equals .09, a result that is not statistically significant ($p = .64$) for a sample of 29, prompting the conclusion that the sensitivity and bias measures are independent.

A second measurement issue explored with the data was the reliability of the measures. Pastore and Scheirer (1974) suggest it is reasonable to assume that the criterion used by a single subject during any measurement session (block of trials) is stable. In other words, in a given set of trials, the response bias should remain stable. A high correlation was expected between the response bias scores obtained in the first and second halves of the recognition test. Also, because stimulus ads in the first and second halves of the portfolio were selected at random, there was no *a priori* reason to believe that overall memory for the ads in the two halves should be different. However, evidence suggests there is an order effect in recognition tasks, in that ads appearing toward the end of the sequence are noted less (Frazen 1942; Lucas and Britt 1963; Starch 1964). Thus, A' was expected to be higher for the first half of the ads than for the second half. Both expectations are supported. Across all subjects, the average correlation between the two B'_H measures is .70 ($p < .01$). The average A' for the first and second halves is .78 and .73, respectively. Further, a comparison of the raw hit rates across all subjects shows

[7]The nonparametric A' measure of sensitivity was suggested by Pollack and Norman (1964) and the nonparametric B'_H measure of response bias was suggested by Hodos (1970).

an 8% decline in recognition scores from .65 for the first half to .57 for the second, which supports the order effect notion.

A third issue explored with the data was the relationship, if any, between the bias measure and the number of magazines read. Multimagazine readers have been found to say "yes, I have seen the ad" more often than persons who read less (Appel and Blum 1961). Therefore, a negative correlation was expected between the number of magazines a person reads and the person's response bias measure, B'_H, if the person is a nay-sayer. The reason is that respondents who are conservative by nature or who are nay-sayers have positive B'_H values. As the number of magazines they "read" increases, so does their propensity to say "yes," thereby lowering their B'_H value. The reverse would hold for yea-sayers. They have negative B'_H values to begin with; an increase in their magazine readership would increase their tendency to say "yes," which would make their B'_H values larger. Though the correlations are in the right direction for both of these expectations, they are not statistically significant. Twenty-two subjects have positive B'_H scores indicating they are nay-sayers. The average correlation between these subjects' B'_H scores and the number of magazines they read is $-.09$ ($p = .70$). For the seven subjects with negative B'_H scores indicating they are yea-sayers, the correlation is .59 ($p = .24$).

The fourth issue addressed with the experimental data was the relation between the sensitivity measure and the average confidence rating across recognition judgments. A number of studies have suggested that as the difficulty of the recognition task increases, as would be the case when the stimulus and distractor items are made more similar, people tend to become less confident of their judgments. Bower and Glass (1976) report, for example, that subjects made significantly more errors and were significantly less confident in their judgments on pairs of items with structurally similar distractors in a forced choice test. Tulving (1981) and Weaver and Stanny (1978) report similar findings. Singh (1982) collected recognition responses of subjects in a 9-alternative forced choice test over six weeks along with their confidence ratings on a 3-point scale—"absolutely confident," "reasonably confident," and "not confident at all." He found that the proportion of responses assigned to the "not confident" category increased from 24% in the first week to 47% in the sixth week, whereas the proportion of responses assigned to the "absolutely confident" category decreased over time from 57% in the first week to 24% in the sixth week. Also, the percentage of wrong responses was higher in the "not confident" category than in the "absolutely confident" category. These studies thus suggest that a subject who is not confident of his or her recognition judgments is more likely to be wrong than right. In other words, there should be a negative correlation between the average confidence ratings when a person's recognition responses are incorrect and the person's measure of recognition memory. The reverse is not true, however; a person who is very confident is not necessarily right in his or her recognition judgments in that the person may be confident for the wrong reasons. For example, a person simply may be very high in generalized self-confidence and may fervently believe his or her recognition judgments are correct even when they are wrong. In sum, our expectations were that (1) there would be a negative correlation between the subject's confidence rating that he or she has made a correct judgment and the person's sensitivity measure A', for those items the subject answered incorrectly and (2) there would be little correlation between the measures for those items the subject answered correctly. Both expectations are confirmed. The correlation between the mean confidence ratings on false alarm responses and the sensitivity measure is $-.31$ ($p < .10$). The correlation between the mean confidence ratings on hit responses and the sensitivity measure is only .01 ($p < .99$).

EXPERIMENT 2

We were encouraged by the results of experiment 1 that addressed issues fundamental to TSD's usefulness in ad recognition testing. The findings that the sensitivity and bias indices were both reliable but independent are important. So are the findings that the response bias measure behaves as expected with respect to magazine readership, at least in a directional sense, and that the sensitivity measure behaves as expected in relation to subjects' confidence in their judgments. A key issue not explored in experiment 1, but which has important implications for ad recognition testing by TSD, is how one chooses the distractor ads for the test.

The issue centers on the notion that performance in a recognition test depends not only on the memory for the stimulus ads but also on the nature and type of distractor ads used in the test. For example, if the subjects were exposed to a set of stimulus ads and later were tested for their recognition memory for those ads by a test in which the stimulus ads were mixed at random with the distractor ads, by merely changing the distractor ads one could obtain a different recognition score even though true recognition memory for the ads should remain the same. The fundamental question is whether adjustment by A' could account for the effect of the distractor ads on the observed recognition scores. Experiment 2 was designed specifically to address this issue.

Method

Three ad portfolios were prepared for experiment 2. Portfolios 1 and 2 were the same as in experiment 1—portfolio 1 contained the 48 stimulus ads and portfolio 2 contained 96 ads—48 stimulus and 48 distractor ads that matched the stimulus ads on a product category basis. Portfolio 3 also contained 48 stimulus and 48 distractor ads. However, the distractor ads for portfolio 3 represented completely different product categories than those used for the stimulus ads. Portfolio 2 could thus be called a "product category congruent stimulus-dis-

tractor portfolio" and portfolio 3 could be called a "product category dissimilar stimulus-distractor portfolio."

For experiment 2, 80 undergraduate students at the same midwestern university were divided randomly into two groups of 40 each. Subjects in both groups saw portfolio 1 and rated each of the stimulus ads as being rational or emotional, just as they did in experiment 1. Subjects were asked to come back later for a second session. Ten subjects from group 1 and two subjects from group 2 did not attend the second session. Subjects in group 1 saw the product congruent portfolio and subjects in group 2 saw the product dissimilar portfolio. The data analysis is based on the responses from the 30 subjects in group 1 and the 38 subjects in group 2 who participated in both sessions.

Results

The four issues investigated in experiment 1 were also investigated in experiment 2 and the results were similar across both groups. Again the sensitivity and bias indices proved to be relatively independent, $r = .31$ ($p = .09$). They also proved to be reliable; the correlation between the response bias scores obtained in the two halves of the recognition test is .61 ($p = .01$). The A' scores for the first and second halves are .81 and .76, respectively. The correlation between the bias measure and magazine readership is again in the expected direction, though again not statistically significant; for the 22 naysayers in the two groups the correlation is $-.32$ ($p = .15$) whereas for the 8 yea-sayers it is .13 ($p = .76$). Finally, the relationship between the sensitivity measure and the average confidence ratings is similar to that for experiment 1. The correlation between the mean confidence ratings and the sensitivity measure is $-.18$ ($p = .35$) for false alarm responses and $-.03$ ($p = .87$) for hit responses.

The differentiating feature of experiment 2 is that it allowed investigation of the impact of choice of distractors. The *a priori* expectation was that it would be more difficult for subjects in group 1 than for those in group 2 to recognize the stimulus ads because of product category congruence between the stimulus and distractor ads. That indeed turned out to be the case. The average raw recognition score across all 48 stimulus ads is 65.4% for group 1 and 77.7% for group 2. Further, the average A' score computed across all subjects in each group is 79.1% for group 1 and 89.7% for group 2, suggesting the recognition task is more difficult when the distractor ads reflect the same products as the stimulus ads. In essence, the manipulation worked.

The fundamental issue that needed addressing was whether the corrections for response bias and discrimination memory worked. Because the portfolios seen by subjects in groups 1 and 2 differed only in the type of distractor ads used, there was no *a priori* reason to believe that memory for ads would differ between the two portfolios. Stimulus ads that are inherently more memorable should be perceived as such by subjects in both groups. This expectation suggests that if the raw recognition scores for each ad were corrected by B'_H to account for the response biases of subjects and by A' to account for the differences in the distractor ads, the *ranking* of the stimulus ads with respect to the adjusted recognition scores should be the same in both groups.

To investigate this question, we adjusted the raw recognition scores for each stimulus ad in each portfolio for B'_H and A'. Next, we rank ordered the ads in descending order from the highest scoring to the lowest scoring ad. That step produced three separate rankings for the stimulus ads in each portfolio, rankings based on (1) unadjusted recognition scores, (2) B'_H-adjusted recognition scores, and (3) B'_H- and A'-adjusted recognition scores. We then calculated Spearman's rank order correlation coefficient for each pair of corresponding ad rankings in the two portfolios. The correlation between the unadjusted recognition scores is $-.45$, between the B'_H-adjusted recognition scores .85, and between the B'_H- and A'-adjusted recognition scores .92.

These correlations demonstrate that the adjustments worked and that the adjusted scores are much more valid than raw recognition scores. The correlation between the ranks based on raw scores is extremely poor. It improves dramatically when the influence of response biases is removed through the B'_H adjustment. It improves still further when the differences in recognition memory due to the differences in distractor ads are removed.

DISCUSSION

Recognition has been a very popular but controversial measure of memory for print advertisements. TSD offers marketers a valuable perspective and useful tool for improving ad recognition tests. The theory suggests that subjects' recognition responses are affected by their response tendencies and their discriminatory abilities. Both of these characteristics can be measured for individual subjects if one has a set of judgments from each subject, some of which are right and some of which are wrong.[8]

To estimate the response bias parameter, B_H, and the discriminatory ability parameter, A', for each subject, a few minor changes will have to be made in the established recognition testing procedures used by such syndicated services as Starch/INRA/Hooper. The ads will have to be removed from the test issue of the magazine and put in a folder along with some number of distractor ads. The subjects will have to be informed about the

[8]Another approach recently described in the marketing literature to account for the response biases and discrimination abilities of subjects makes use of the beta binomial probability model. It allows the estimation of the *distribution* of true discrimination ability over subjects (Schmittlein 1984; Schmittlein and Morrison 1983), but does not provide individual subject measures for response bias and discrimination ability. The subject-by-subject measures seem to be very useful to advertisers because they allow the adjustment of recognition of persons claiming to have seen the advertiser's product, which is, of course, the one in which the advertiser is most interested.

presence and number of distractor ads in the test. Removing the test ads from the context of the magazine has at least two advantages. First, it eliminates the possibility of falsely recognizing an ad from the context of the reading material. Second, the position of the test ads can be rotated in a portfolio so that there is no order effect (Lucas and Britt 1963).

Given the estimates B'_H and A', several adjustments can be made to the raw scores to make them more meaningful. Consider first an adjustment for B'_H. Because B'_H is a measure of how prone a subject is to saying "yes"—independent of his or her memory for the ad—B'_H can be used for obtaining an index of recognition memory for each individual that is adjusted for response biases. Recall that B'_H is expressed as a percentage and that a higher B'_H indicates a lower tendency for saying "yes." Therefore, given equal memory across two subjects for the ads in a certain recognition test, a subject with $B'_H = -20\%$ is twice as prone to say "yes" on a given recognition test as a subject with $B'_H = -10\%$. Further, though B'_H is computed across all ads presented in a test session, it can be used to ascertain recognition scores adjusted for response tendencies for individual ads because it is reasonable to assume that the decision criterion (B'_H) employed by a single subject in a given measurement is stable (Pastore and Scheirer 1974), that is, the response bias remains constant.

The adjustment is easy to make. The only requirement is that B'_H be computed for each subject in the test from the person's hit and false alarm rates. The responses of subjects claiming "yes, I have seen a particular ad" would then be coded 1 and "no" responses would be coded 0. The dummy variables would be multiplied by the B'_H values per subject, the products would be summed, and the sums would be divided by the number of subjects to generate an average adjusted score per ad. These scores would be response bias adjusted scores.

The response bias adjusted recognition indices seem very useful for comparing recognition scores across samples in addition to their use in comparing recognition scores for various ads within a given sample of subjects. For example, a manager may be interested in knowing whether an ad appearing in both the April and May issues of a monthly magazine registered better recognition in one month than in the other. Because normally two different samples of subjects would be used to obtain the recognition scores for the ads appearing in the two different issues, it is very likely that the samples would differ in their response biases. Hence, a comparison of unadjusted recognition scores may be misleading and the comparison of adjusted recognition indices would be more appropriate.

Consider next the additional adjustment for A' to reflect the notion that the recognition scores obtained in a typical recognition test depend not only on the subject's memory for stimulus ads and his or her response biases, but also on the similarity between the stimulus and distractor items used in the test. The more similar the distractor ads are to the stimulus ads, the more difficult would be the discrimination task and the lower the recognition scores. The A' adjustment accounts for the similarity of the ads. The procedure one would use to generate this adjustment would parallel that used to generate the response tendency adjusted index. More specifically, the dummy variables reflecting the yes/no responses would be multiplied by B'_H and A' scores by subject, the products would be summed, and the average value across subjects computed. The index formed thus would logically be called a "global index" because it adjusts for both response tendencies of subjects and similarity of ads. Though the global index could be used to compare recognition scores across subjects in a single testing session, its most productive use would be in adjusting the scores obtained across tests to sort out the impact of different degrees of similarity between stimulus and distractor ads as well as the inherent changing composition of the samples of subjects used to secure recognition scores.

All the adjustments mentioned reflect basic subject differences. However, the indices are also capable of adjusting raw recognition scores for externally induced contaminants affecting a subject's responses. For example, the differences in the response tendencies induced by interviewer differences in asking questions, or in providing clear or ambiguous instructions, could be adjusted by these indices. If the interviewers by their behaviors caused respondents to increase in their own minds the psychic value of a hit versus a correct rejection, this effect would show up in B'_H-adjusted scores.

TSD also could be applied to recognition testing of broadcast ads. For example, using self-administered questionaires, Bruzzonne Inc., the only company that performs such tests, collects recognition information on a number of broadcast ads. Because these tests yield hit rates and false alarm rates for individual subjects, TSD could be applied to Bruzzonne data without any modifications in their testing procedure.

One criticism that almost surely will be made of the adjusted recognition indices is that they are not nearly as interpretable as raw recognition scores. In one way that argument makes sense. An unadjusted recognition score has great intuitive appeal; the statement that 30% of the sample group remembered seeing the ad can be easily understood by the least sophisticated manager. The statement that the ad's response bias adjusted recognition score is .8, say, is not as easily interpreted. The problem, though, does not represent a fundamental deficiency in the index, just limited experience in using it. As researchers gather experience with these indices, they can begin to generate distributions reflecting the frequency with which the various values occur. By referencing subsequent values of the indices to the distributions, one would have an equally interpretable measure; thus one could talk about the fact that, say, the empirical evidence suggests a particular adjusted index value occurs less than 70% of the time, suggesting the ad gen-

erated a high level of recognition memory. This conclusion certainly seems preferable to arguing that 70% of the people claimed they saw the ad, but the number who really saw it is not known because of false claiming tendencies. In sum, norms that reflect the frequency with which each value occurs could be developed for each index. The norms could be developed for specific types of products as well as different media and different data collection processes. In each case, referencing the raw data to the appropriate norms would provide a much truer indication of the recognition memory of an ad.

REFERENCES

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

Banks, W. P. (1970), "Signal Detection Theory and Human Memory," *Psychological Bulletin*, 74 (August), 81–99.

Bower, G. H. and A. L. Glass (1976), "Structural Limits and the Reintegrative Power of Picture Fragments," *Journal of Experimental Psychology: Human Learning and Memory*, 2 (July), 456–66.

Clancy, K. J., L. E. Ostlund, and G. A. Wyner (1979), "False Reporting of Magazine Readership," *Journal of Advertising Research*, 19 (October), 23–30.

Coombs, C. H., R. M. Dawes, and A. Tversky (1970), *Mathematical Psychology*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Corso, J. F. (1967), *The Experimental Psychology of Sensory Behavior*. New York: Holt, Rinehart and Winston, Inc.

Cronbach, L. J. (1950), "Further Evidence on Response Sets and Test Design," *Educational and Psychological Measurement*, 10 (Spring), 3–31.

Davenport, J. S., E. B. Parker, and S. A. Smith (1962), "Measuring Readership of Newspaper Advertisements," *Journal of Advertising Research*, 2 (December), 2–9.

Egan, J. P. and F. R. Clarke (1966), "Psychophysics and Signal Detection," in J. B. Sidowski, ed. *Experimental Methods and Instrumentation in Psychology*. New York: McGraw-Hill Book Company.

Elliott, P. B. (1964), "Tables of d′," in *Signal Detection and Recognition by Human Observers*, J. A. Swets, ed. New York: John Wiley & Sons, Inc.

Frazen, R. (1942), "Inequalities Which Affect Scores on Advertisements," *Journal of Marketing*, 6 (April), 128–32.

Glass, A. L., K. J. Holyoak, and J. L. Santa (1979), *Cognition*. Reading, MA: Addison-Wesley Publishing Company.

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

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

Guilford, J. P. (1967), "Response Biases and Response Sets," in *Readings in Attitude Theory and Measurement*, Martin Fishbein, ed. New York: John Wiley & Sons, Inc.

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

Klatzky, R. L. (1980), *Human Memory: Structures and Processes*. San Francisco: W. H. Freeman & Company.

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

——— and S. H. Britt (1963), *Measuring Advertising Effectiveness*. New York: McGraw-Hill Book Company.

Marder, Eric and M. David (1961), "Recognition of Ad Elements: Recall or Projection?," *Journal of Advertising Research*, 1 (December), 23–5.

Moran, W. T. (1951a), "Measuring Exposure to Advertisements," *Journal of Applied Psychology*, 35 (February), 72–7.

——— (1951b), "A Reply to Heller's Note," *Journal of Applied Psychology*, 35 (February), 78–9.

Neu, D. M. (1961), "Measuring Advertisement Recognition," *Journal of Advertising Research*, 1 (December), 17–22.

Pastore, R. E. and C. J. Scheirer (1974), "Signal Detection Theory: Considerations for General Application," *Psychological Bulletin*, 81 (December), 945–58.

Peterson, W. W., T. G. Birdsall, and W. C. Fox (1954), "The Theory of Signal Detectability," *Transactions IRE Professional Group on Information Theory*, 4 (September), 171–212.

Pollack, S. and D. A. Norman (1964), "A Nonparametric Analysis of Recognition Experiments," *Psychonomic Science*, 1 (May), 125–6.

Schmittlein, D. C. (1984), "Assessing Validity and Test-Retest Reliability for 'Pick K of N' Data," *Marketing Science*, 3 (Winter), 23–40.

——— and D. G. Morrison (1983), "Measuring Miscomprehension for Televised Communication Using True-False Questions," *Journal of Consumer Research*, 10 (September), 147–56.

Simmons, W. R. (1961), "Controlled Recognition in the Measurement of Advertising Perception," *Public Opinion Quarterly*, 25 (Fall), 470–71 (abstract).

Singh, S. N. (1982), *Recognition as a Measure of Learning from Television Commercials*, unpublished doctoral dissertation. Ann Arbor, MI: University Microfilms International.

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

Starch, D. (1946), *Factors in Readership Measurement*. New York: Daniel Starch and Staff.

Swets, J. A., W. P. Tanner, Jr., and T. G. Birdsall (1964), "Decision Processes in Perception," in *Signal Detection and Recognition by Human Observers*, J. A. Swets, ed. New York: John Wiley & Sons, Inc.

Tulving, E. (1981), "Similarity Relations in Recognition," *Journal of Verbal Learning and Verbal Behavior*, 20 (October), 479–96.

Van Meter, D. and D. Middleton (1954), "Modern Statistical Approaches to Reception in Communication Theory," *Transactions IRE Professional Group on Information Theory*, 4 (September), 119–41.

Wallace, W. P. (1980), "False Recognition Produced by Implicit Verbal Responses," *Psychological Bulletin*, 88 (November), 686–704.

Weaver, G. E. and C. J. Stanny (1978), "Short Term Retention of Pictorial Stimuli as Accessed by a Probe Recognition Technique," *Journal of Experimental Psychology: Human Learning and Memory*, 4 (January), 55–65.

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