ALAN G. SAWYER and J. PAUL PETER*

Classical statistical significance testing is the primary method by which marketing researchers empirically test hypotheses and draw inferences about theories. The authors discuss the interpretation and value of classical statistical significance tests and suggest that classical inferential statistics may be misinterpreted and overvalued by marketing researchers in judging research results. Replication, Bayesian hypothesis testing, meta-analysis, and strong inference are examined as approaches for augmenting conventional statistical analyses.

The Significance of Statistical Significance Tests in Marketing Research

Two basic types of empirical evidence used in hypothesis testing in marketing research are observations of covariation and observations of differences between groups. This evidence usually consists of sample data, and the acceptability of the evidence is based almost inevitably on classical statistical significance tests. However, a review of marketing research texts and a variety of marketing research articles leads to the conclusion that, in both theory and practice, the logic of statistical significance testing is sometimes misinterpreted in the marketing literature. Perhaps because of this misinterpretation, marketing researchers may seriously overvalue the role of classical inferential statistics in the research process.

The purpose of our article is to examine the interpretation and value of statistical significance testing and to offer recommendations to improve the quality of hypothesis testing in marketing research. Although the issues we discuss pertain directly to data from experimental research, most of these issues also apply to correlational and other data. Many examples, including several from the marketing literature, illustrate our recommendations. We hope to persuade more marketing researchers to follow their lead. Though we are not arguing against using classical inferential statistics for what they were designed to do, we are concerned with the tendencies to endow them with capabilities they do not have and to utilize them as the sole approach to analyzing research data.

These problems are not exclusive to research in marketing. Writers in psychology (e.g., Bakan 1966; Lykken 1968), sociology (e.g., Henkel 1976; Selvin 1957), and education (e.g., Carver 1978) have argued that these misinterpretations also pervade their disciplines. In fact, tests of statistical significance seem to be relied upon and often misused in all the social sciences. In comparison, perhaps because of the more highly developed theory, more reliable measurement techniques, and greater opportunity to control nuisance variables, researchers in the physical sciences often forego inferential statistical tests and instead focus directly on the data themselves. Although marketing phenomena may not lend themselves to this approach (Peter 1983) and the theory, measurement, and research procedures in marketing may never develop sufficiently for us to follow the analytical practices of the physical sciences, researchers should be more aware of the limitations of most inferential statistics and the value of augmenting them with other information and other research approaches.

INTERPRETATION OF STATISTICAL SIGNIFICANCE TESTS

For a proper interpretation of the meaning of a statistically significant result, the assumptions of the classical statistical significance testing model must be understood. A primary assumption is that the null hypothesis (e.g., no difference between treatment effects, no association between variables) is true and any observed differences

*Alan G Sawyer is Associate Professor of Marketing, The Ohio State University J Paul Peter is Associate Professor of Marketing, University of Wisconsin–Madison

The authors gratefully acknowledge the time and financial support from the Dean's Research Fund of the College of Administrative Science at The Ohio State University and from The School of Business, University of Wisconsin–Madison, as well as the helpful comments of several colleagues including Peter Dickson, Jim Ginter, Mike Houston, Glenn Milligan, and anonymous *JMR* reviewers

Journal of Marketing Research
Vol XX (May 1983), 122–33

or associations are the result of sampling error. For example, a statistically significant mean difference at $p \leq .05$ tells us that if we sampled many pairs of groups from the same hypothetical population, we would expect to get a difference as large as the observed result or larger with no more than 5% of the groups as the result of sampling error, given that the null hypothesis is true. In general terms, a statistically significant result is one which occurs rarely if the null hypothesis is true.

Many writers in social science have commented on the failure of researchers and textbook writers to interpret statistical significance correctly. In a recent summary of much of this work, Carver (1978) discusses three common misinterpretations, all three of which can be found in the marketing literature. These three misinterpretations are that a statistically significant result indicates (1) the probability that the results occurred because of chance, (2) the probability that the results will be replicated in the future, and (3) the probability that the alternative hypothesis is true. A fourth misinterpretation involves confusion about the role of sample size and the level of statistical significance.

The Probability of the Null Hypothesis

The first misinterpretation is to view a p-value as the probability that the results occurred because of sampling error or chance fluctuations. For example, $p = .05$ is interpreted to mean that there is a probability of only .05 that the results were caused by chance. However, this interpretation is completely erroneous because (1) the p-value was calculated by assuming that the probability is 1.0 that any differences were the result of chance and (2) the p-value is used to decide whether to accept or reject the idea that the probability is 1.0 that chance caused the mean difference. A p-value of .05 means that, *if* the null hypothesis is true, the odds are 1 in 20 of getting a mean difference this large or larger and the odds are 19 in 20 of getting a smaller mean difference. However, *there is no way in classical statistical significance testing to determine whether the null hypothesis is true or the probability that it is true*. As Cronbach and Snow (1977, p. 52) explain:

> A p value reached by classical methods is not a summary of the data. Nor does the p value attached to a result tell how strong or dependable the particular result is . . .. Writers and readers are all too likely to read .05 as p(H/E), "the probability that the Hypothesis is true, given the Evidence." As textbooks on statistics reiterate almost in vain, p is p(E/H), the probability that this Evidence would arise if the (null) hypothesis is true. Only Bayesian statistics yield statements about p(H/E).

The Probability of Results Being Replicated

A second misinterpretation is that the p-value represents the confidence a researcher can have that a given result is reliable or can be replicated. Basically, this argument is that the complement of the p-value yields the probability that a result is replicable or reliable, e.g., 1 − .05 = .95 probability that results can be replicated. This misinterpretation probably comes from a notion that a statistically significant difference in sample means suggests that the samples are from different hypothetical populations and future samples drawn from these different hypothetical populations will therefore yield equivalent results. However, *nothing in classical statistical significance testing says anything about the probability that the same results will occur in future studies*. Replicating results is a function of how exactly the method is repeated, and some aspects, such as the time of measurement, clearly cannot be identical to those of the original study.

The Probability of Results Being Valid

The third and most serious misinterpretation of classical statistical significance testing is that it directly assesses the probability that the research (alternative) hypothesis is true. For example, a p-value of .05 is interpreted to mean that its complement, .95, is the probability that the research hypothesis is true. Related to this misinterpretation is the practice of interpreting p-values as a measure of the degree of validity of research results, i.e., a p-value such as $p < .0001$ is "highly statistically significant" or "highly significant" and therefore much more valid than a p-value of, say, .05. Again, such a practice is inappropriate. Although it is true, for example, that the greater the difference between group means the greater the chance of obtaining a small p-value, and it is true that such a result may be rarer given the null hypothesis, a statistically significant result cannot properly be construed as a more valid result for at least two reasons.

First, a statistical test is not a complete test of a research hypothesis. Instead it examines only one of many possible operationalizations of a research hypothesis. Thus, it is improper to infer that the research hypothesis is valid without testing and support from a representative sample of operationalizations. Second, a variety of threats to drawing valid inferences are not addressed by statistical tests (Cook and Campbell 1979). Theoretically, the researcher's job is to eliminate or at least to render implausible all of the alternative explanations before accepting the research hypothesis. However, this task is no small matter given the variety of theories and method factors that can be offered to explain any empirical result. When these variables along with possible higher order interactions are considered, the task becomes even more difficult (see Cronbach and Snow 1977). In any event, *rejection of the null hypothesis at a predetermined p-level supports the inference that sampling error is an unlikely explanation of results but gives no direct evidence that the alternative hypothesis is valid.*

Sample Size and the Probability of the Research Hypothesis

A fourth common misinterpretation about statistical testing involves the relationship between sample size and

level of statistical significance. If a given relationship is found to be statistically significant at a given confidence level, it is sometimes implied that more confidence should accompany this result if the study had a large sample size rather than a small one. Rosenthal and Gaito (1963) report that such a conclusion was very prevalent among the research psychologists they surveyed. However, such a conclusion is false. Larger samples do reduce likely sampling error because their estimates more closely approximate the population parameters, but it should also be clear that differences in the amount of sampling error are included explicitly in the computation of statistical significance tests. Thus, there should not be a bias against statistically significant results obtained from properly selected small samples.

Moreover, because effect size is a measure of the strength of the relationship and large effects are more likely to be replicated than small ones, researchers should have more confidence in the study with the smaller sample. Meyer (1974) demonstrates this fact with a Bayesian analysis of binominal data with results for different sample sizes. Meyer's results stem from the simple fact that smaller effect sizes are considered statistically significant with larger sample sizes and that, though a larger sample size helps to reduce sampling error and the resulting higher statistical power of a classical inferential statistic increases the probability of a rejection of the null hypothesis, it does not necessarily increase the probability of a *valid* rejection.

THE VALUE OF STATISTICAL SIGNIFICANCE TESTS

The value of statistical significance testing is severely restricted because it does not accomplish what researchers often want or, perhaps in some cases, assume that it does. Several factors detract from the value of statistical tests. First, the process of statistical hypothesis testing is hardly objective given the many subjective decisions made by the researcher. Second, exact null hypotheses are very rarely true in the population, and researchers typically are very biased against the null hypothesis in their testing procedures. Third, classical statistical significance tests are often uninformative without various descriptive statistics and other inferential tests such as confidence intervals. Finally, classical statistics offer no direct evidence about individual behavior.

The Subjectivity of Statistical Tests

Perhaps a major factor contributing to the perceived value of statistical significance tests is the illusion that they are completely objective. Though such tests are mathematical and precise[1] and provide "a formal and nonsubjective way of deciding whether a given set of data shows haphazard or systematic variation" (Winch and Campbell 1969, p. 143), one should not infer that they are objective tests. The reason is that whether a given statistical significance level is obtained is strongly influenced by subjective decisions by the researcher. As Bakan (1966, p. 426) points out, "the probability of rejecting the null hypothesis is a function of five factors: Whether the test is one or two-tailed, the level of significance, the standard deviation, the amount of deviation from the null hypothesis and the number of observations." The researcher clearly controls the first, second, and fifth factors and can influence the third and fourth. Thus, many obtained results which are not statistically significant can become so by such methods as (1) increasing the sample size, (2) increasing the reliability of the measures, (3) changing *post hoc* the acceptable level of statistical significance (i.e., from .01 to .05), (4) changing from a two-tailed to a one-tailed test, and (5) obtaining better control over nonmanipulated variables. *Because researchers make many subjective decisions that greatly influence the probability of rejecting the null hypothesis, it is misleading to consider the process of statistical significance testing as objective solely because of the objectivity of the mathematics.*

A methodological paradox in social science research relates to the illusion of objectivity (Meehl 1967). Methodological improvements such as increased control, more precise measurement, and a greater number of observations make it *easier* for the social scientist to reject the null hypothesis (and claim support for the alternative hypothesis), whereas such improvements make it more difficult for the physical scientist to reject the null hypothesis. The reason for this paradox is that, in the physical sciences, theory is often used to predict point values and, if used at all, statistical significance tests evaluate the difference between the value predicted by the theory and the value found in the data. In contrast, most social science theories are not developed sufficiently to make point predictions and instead statistical significance tests are used to test all other values against the null hypothesis of zero. Meehl suggests that the use of statistical significance testing in social science thus makes it very difficult to not reject the null hypothesis and the involved theory.

Research Bias Against the Null Hypothesis

Classical statistical significance tests are set up under the assumption that the null hypothesis is true. Such an assumption is, in fact, almost always false, and much well intended marketing research practice is biased against the null hypothesis. First, null hypotheses of no treatment effect or no relationship are almost always false because, in the population, few behavioral variables ever

[1]Some marketing researchers have tried to quantify problems with collected data other than sampling errors (e g , Brown 1967, Lipstein 1975, Mayer 1970) Though not optimistic about the ability to quantify these many other types of errors, we applaud the effort because it helps point out the obvious limitations of a statistic that precisely quantifies what very often is the least serious of the many threats to accurate estimation (see, for example, Assael and Keon 1982)

have *exactly* a zero mean difference between two groups or an exactly zero correlation with each other. For example, Meehl (1967) reported that 91% of pairwise associations among 45 variables in a sample of 55,000 people were statistically significant, and Bakan (1966) failed to find any statistically insignificant relationships among many tests in a sample of 60,000. Given sufficiently high statistical power, one would expect virtually *always* to conclude that the exact null hypothesis is false. It is no wonder that "statistical significance" has occurred often in recently published marketing research because these studies typically have relatively high statistical power (Sawyer and Ball 1981). We find it frightening to consider how much of the conventional wisdom in marketing is based on little evidence other than statistical significance.

Researchers and publication practices are biased against the null hypothesis. Researchers inevitably expect to reject the null, and publication practices overwhelmingly favor studies which achieve this objective. Greenwald (1975a) describes how researchers are unlikely even to try to publish results of an empirical study that failed to reject the null, and journals are even less likely to accept the few statistically insignificant-result studies that are submitted. In an extensive review, Glass, McGaw, and Smith (1981) determined that "findings reported in journals are, on the average, one-third standard deviation more disposed toward the favored hypothesis of the investigators than findings reported in theses and dissertations" (p. 67).

Such a selection bias toward submitting and/or publishing only statistically significant results leads to the fear that "file drawers" are full of statistically insignificant studies and that the published ones are the only ones that attain conventional statistical significance. Using measures of effect size, Rosenthal (1979) demonstrates how to incorporate the possibility of "file drawer" support for the null hypothesis into calculations of possible Type I error and concludes that, "when the number of studies available grows large or the mean directional Z (effect size) grows large, the file drawer hypothesis as a plausible rival hypothesis can be safely ruled out" (p. 640). In contrast, with a small sample of statistically significant studies, relatively few "filed" studies with "insignificant" results would have to exist to yield a net statistically insignificant conclusion. For example, according to Rosenthal, 15 studies with an average effect size of $Z = .50$ have a combined Type I error rate of $p = .026$, but, if there were as few as six other studies showing a mean effect size of .00, the overall set of results would be judged statistically insignificant (i.e., $p > .05$).

After one rechecks the calculations (Rosenthal 1969),[2] the typical reaction to a failure to reject a null hypothesis is to blame the failure on something wrong such as a weak manipulation, a small sample size, or unreliable measurement (McGuire 1973). Even when several failures to reject a null hypothesis are reported in the literature, researchers still cling to the alternative hypothesis as the most likely (e.g., Cartwright 1973). Our suspicion that marketing researchers suffer from a similar bias is based on our inability to recall any instances in which it is widely agreed that a previously hypothesized relationship does not hold. Apparently, *results from statistical significance tests are perceived to be valuable when they support the favored hypothesis but are commonly discounted when they support the null.*

The Need for Descriptive Statistics

A major problem in the use and reporting of classical statistical significance tests is that they commonly appear to dominate or even substitute for the data themselves. Frequently, tables of F-values are discussed before or instead of such vital descriptive statistics as means and confidence intervals. Such priority is clearly misinforming as well as misinformed. *The major results of any empirical study, regardless of whether the prime purpose is description, prediction, or explanation, are the descriptive statistics that indicate the nature and size of any obtained effects.* As Sawyer and Ball (1981) summarize, statistical significance tests do not say anything about the size or importance of an effect. Lower Type I error probabilities do not necessarily imply a larger effect. A very small effect can be statistically significant with a sufficiently large sample; conversely, a sizable effect can be judged statistically insignificant with a very small sample. Effect size can be measured in many ways including R^2, ω^2, and other estimates of the ratio of explained to total variance; alternatively, various expressions of the standardized mean difference between groups such as Z or Cohen's (1977) d values can be used (see also Rosenthal and Rubin 1982).

Statistics and Individual Behavior

A final point that is occasionally overlooked about the value of statistical significance tests is that they focus on aggregate central tendencies and reflect little about individual behavior. One interesting way to illustrate this point is to consider Cohen's (1977) U descriptive statistic which measures the percentage overlap between two distributions. Even with a reasonably large effect size, a large percentage of individuals in two groups will often be essentially similar or ordered contrary to the direction of the overall group mean. For example, Cohen states that a difference as large as .8 of a standard deviation is relatively "large" for much social science research. Even

[2]Lest the reader doubt this, we ask the following question After having calculated, for example, an F-value that suggests your favored research hypothesis is statistically significant, how likely are you to recheck your figures, make sure your computer format statement was correct, etc ? Alternatively, how many hours have you spent checking and rechecking data that failed to attain statistical significance? Interestingly, Rosenthal (1969) observed that when computational errors occur, nearly three-fourths of those errors are in the direction of the researcher's hypothesis.

with such a difference, however, 52.6% of the two populations are overlapped. Thus, though marketing researchers frequently conclude that, for example, new product adopters are different from nonadopters in a certain way, it is almost always erroneous to conclude that *all* adopters are different from *all* nonadopters in that way and, in most instances, wrong to infer even that *most* adopters are different in a given way. Although this is often the type of conclusion researchers want to draw, a statistical significance test alone does not justify such a conclusion.[3]

RECOMMENDATIONS

We offer several recommendations designed to address the problems discussed and to strengthen hypothesis testing in marketing. First we present three considerations for improving the use of classical statistical significance tests against the null hypothesis. We then describe and illustrate four research perspectives that provide valuable additional information about research questions: replication, Bayesian hypothesis testing, meta-analysis, and strong inference.

Tests Against the Null Hypothesis

We do not recommend as do some writers (e.g., Carver 1978; Henkel 1976) that classical statistical significance testing be discarded. Statistical significance testing is a useful "act of discipline" (Cronbach and Snow 1977) to sort out findings that may be worthy of more attention. However, marketing researchers should become more aware of the limited value of classical statistical significance tests. We offer three recommendations for improved practice in the use of classical statistical tests against the null hypothesis.

First, we support Kish's (1959) recommendation of two decades ago that the phrase "test against the null hypothesis" be substituted for the ambiguous and potentially misleading phrase "test of significance" to avoid miscommunication about the proper meaning of statistical tests. Furthermore, though results may be "statistically significant" they should not be reported as "significant" or "highly significant" which suggests that they are valid or important or provide a measure of effect size. Researchers also should avoid the misleading impression of precision or objectivity by reporting the exact statistical significance level to the fourth decimal place.

Second, because point null hypotheses are of limited value, a range rather than a point null hypothesis should be employed if possible. A range null hypothesis requires a decision in advance of data collection about the lowest effect size that will be considered to be of consequence or nontrivial. Any result within the range of effects smaller than the specified minimum would be judged as evidence that fails to reject the null hypothesis. Even if the decision is an arbitrary one, such a practice can lead to more meaningful use of tests against the null hypothesis because the range constituting the null hypothesis then becomes a respected alternative instead of a "straw man." At the very least, point null hypotheses should be replaced by a directional hypothesis; a theory that cannot generate at least a directional prediction is unworthy of the term "theory." As Meehl (1978, p. 825) forcefully argues, "It is always more valuable to show approximate agreement of observations with a theoretically predicted numerical point value, rank order, or function form, than it is to compute a 'precise probability' that something merely differs from something else."

By recommending use of directional hypotheses, we simply mean that investigators should make their expectations explicit to both themselves and others instead of following the traditional practice of stating hypotheses in the null form. However, we do not want to appear to support the practice of using one-tailed tests to prove that, for example, a t-value of 1.69 is "significant." Such emphasis on p-values gives them undue importance and diverts attention from effect size estimates. Furthermore, the tentativeness of any marketing theory ought to be recognized explicitly by more conservative two-tailed statistical tests.

Third and most important, empirical results should be described and analyzed such that the size and substantive significance of obtained effects are emphasized and not merely the p-values associated with the resulting test statistics. Presenting appropriate descriptive statistics such as means, variances, confidence intervals, contrast estimates, and estimates of total variance accounted for by a given variable before any inferential statistics can help achieve the goal of a more complete description of results. Estimates of the power of an employed statistical test to detect an effect of a chosen size can help the reader to understand more fully the nature of the obtained results and to judge the precision of the chosen inferential statistical test. Reporting statistical power is especially important when the statistical analysis does not reject the null hypothesis (Sawyer and Ball 1981).

Replication

More value should be placed on replication in marketing research. We stated before that statistical significance testing does not provide evidence about the replicability of obtained results. Science, however, depends on replication (cf. Lykken 1968; Smith 1970). If a result is replicated sufficiently, statistical significance tests are unimportant. As Stevens (1971, p. 440) stated:

> In the long run, scientists tend to believe only those results that they can reproduce. There appears to be no better option than to await the outcome of replications. It is probably fair to say that statistical tests of significance, as they are so often miscalled, have never convinced a scientist of anything.

[3]Perhaps more value would be placed on the insights from studies of individual behavior (e.g , Bettman 1974; Krugman 1971) if marketing researchers were concerned less with statistical inference tests than with the data themselves and descriptions of them

Tversky and Kahneman's (1971) results indicate that research scientists are overly confident about the future replicability of a research result which favors the alternative hypothesis. Brown and Gaulden (1980) and Leone and Schultz (1980) have cited the dearth of replication in marketing research. Perhaps our field would not hold replication in such low regard if we were properly less naïve and smug about the interpretation and value of tests against the null hypothesis.

In early stages of research on a given set of hypotheses, replications which come as close as possible to the original study can be valuable for determining the nature and extent of effects. However, even more valuable as well as more efficient than *exact* replications are *balanced* replications. Balanced replications combine exact replications as control conditions with other conditions which manipulate additional substantive and/or methodological variables (see Carlsmith, Ellsworth, and Aronson 1976).

In several recent marketing studies researchers have used replication and statistical analysis of survey data in a manner similar to several of our recommendations. Dodson, Tybout, and Sternthal (1978) used economic utility and self-perception theories to predict and test a series of hypotheses about brand switching after purchasers either used a media-distributed coupon, bought during a cents-off deal, or redeemed a cents-off package coupon. Self-perception theory made successful (and unintuitive) predictions that repeat purchase probability would decrease, not increase, after a purchase with a media-distributed coupon. Results were replicated successfully over two product classes. Although the authors carefully conducted statistical tests of the data, they properly placed emphasis on the data and effect magnitudes.

Bagozzi (1978) similarly used theory from a variety of sources to generate several specific hypotheses about salesforce performance and satisfaction. Bagozzi carefully replicated his results across test and validation subsamples of two different samples of salespeople which differed in terms of experience and need for planning and motivation. The analysis also properly emphasized estimates of effect size such as beta coefficients and R^2. Ryans and Weinberg's (1979) analysis of determinants of territory sales response shares many of the aforementioned qualities, as do Della Bitta, Monroe, and McGinnis' (1981) replicated experiments about different ways to advertise a price reduction. Finally, Eskin and Baron's (1977) series of replicated field experiments which factorially manipulated both price and advertising expenditures is an excellent example of how replications can strengthen confidence in the external validity of a given result—especially when the result is unanticipated such as the price-advertising interaction effect they found in three of four experiments. Eskin (personal communication, 1982) has recently gathered information on about 40 experiments with retail advertising and pricing that further replicate the results of Eskin and Baron.

Bayesian Hypothesis Testing

In applied problems, when replications are not possible before a decision must be made, the use of Bayesian statistics instead of classical statistics is highly advisable. However, Bayesian statistics ought not to be confined to applied decision problems. Bayesian analysis affords several advantages in theoretical research that may not be appreciated by many marketing researchers.

Unlike classical statistical significance testing, the Bayesian approach does estimate a continuous likelihood of p(H/E) and does not necessitate a dichotomous decision that the null hypothesis is either completely false or true (Edwards, Lindman, and Savage 1963; Iverson 1970). The Bayesian approach directly compares the null and alternative hypotheses and allows one to consider more fully the possibility that the null hypothesis is true. Because the posterior distribution may be influenced by the subjective prior probabilities of an individual researcher, some researchers may reject Bayesian statistics for scientific analyses of theoretical propositions. However, as discussed before, classical statistical tests are not free from subjective decisions that can influence results. Bayesian statistics at least force the researcher to specify clearly in the prior distribution any subjectivity that enters the analysis, and allow a determination of the effects of subjective choices on the final conclusions (Iverson 1970). Furthermore, the subjective nature of prior probability estimates can be reduced by adopting a prior distribution which is essentially "flat" or insensitive in the most likely region of effect and which does not favor one extreme over another (Phillips 1973).

Greenwald (1975a,b,c) has demonstrated the greater flexibility of Bayesian hypothesis testing for making a decision between two relevant and feasible hypotheses in theoretical research, and how the Bayesian approach can provide more useful information than classical statistical significance tests when one is analyzing a series of replications. Greenwald (1975c) cited as one example the research of Layton and Turnbull (1975), who conducted two nearly identical experiments which manipulated two independent variables. They found only one small main effect in the first experiment and no statistically significant effects in the second experiment. Layton and Turnbull concluded that, given the results, they were "left with no alternative but to consider these studies *inconclusive* regarding the effects of the experimental manipulations" (p. 178).

Greenwald disputed Layton and Turnbull's conclusion and suggested that reliance on classical statistical tests was to blame for their failure to conclude something from the data of more than 400 subjects in two well-conducted experiments. In his Bayesian reanalysis, Greenwald first defined the minimum effect sizes that the experiments were able to detect. Then, for the first experiment, he formulated a flat prior probability distribution that was not subjectively biased in favor of either the null or alternative hypotheses. He next computed a likelihood

function and a posterior probability distribution for each effect from the data and tested each of the hypotheses in terms of the odds computed from the posterior distribution. The same analysis was performed on the data from the second experiment except that the posterior distributions from the first experiment were used as the priors for the second analysis. The final posterior odds *in favor of the null hypothesis* were 7.8 to 1 for one independent variable and 23.3 to 1 for the other. Greenwald thus concluded that the chances of obtaining results supportive of the alternative hypotheses for either effect were very low. Whereas the original classical statistical analysis resulted in a decision that the findings were inconclusive, Greenwald's Bayesian statistical analysis led to a more definitive conclusion that the effects of the variables in question were likely very small and that, if one wanted to test the likelihood of a null hypothesis, it was much more probable than the alternative.

Unfortunately, only a few published studies in marketing research have employed Bayesian statistics to test hypotheses. An excellent recent example of the advantage of the Bayesian over the classical approach in applied marketing research is discussed by Blattberg (1979) and Ginter et al. (1981). Banks (1965) also gives an extensive example (which was taken from Schlaifer's 1961 textbook), as does Roberts (1963). One exception to the non-utilization of Bayesian hypothesis testing in marketing is Levitt's (1972) reanalysis of his hypotheses about source credibility in industrial selling with Bayesian statistics. Levitt's Bayesian analysis helped to describe better the experimental results without the typical marketing research use of an insignificant classical statistical test as a barrier to examining the data for any valuable information (Zeisel 1955). More marketing researchers ought to use the Bayesian approach.

Meta-Analysis

Researchers' undue reliance on classical statistical tests is illustrated in many literature reviews. Traditional literature reviews often focus on counting the number of studies in a given area which do and do not find statistically significant relationships or differences. However, this approach ignores many of the issues we have raised and can result in misleading conclusions. As Meehl (1978) states, "When a reviewer tries to 'make theoretical sense' out of such a table of favorable and adverse significance test results, what the reviewer is actually engaged in, willy-nilly or unwittingly, is meaningless substantive constructions on the properties of the statistical power function, and almost nothing else" (p. 823). An alternative approach for summarizing previous empirical studies is *meta-analysis* (Glass, McGaw, and Smith 1981; Houston, Peter, and Sawyer 1983).

Meta-analysis involves a quantitative review of a research question and focuses on the obtained effect sizes in previous studies on the topic. In a meta-analysis one attempts to obtain all previous empirical studies pertaining to the research question, including if possible both published and unpublished work. The researcher using meta-analysis seeks general conclusions while searching for methodological conditions and substantive variables that might measurably moderate any observed main effects. To the extent that included studies are of varied quality, study characteristics ought to be coded as well as possible so that the size and direction of any effects of study quality can be assessed in the meta-analysis. A variety of quantitative criteria (including statistical tests) have been suggested for summarizing results. However, Glass, McGaw, and Smith (1981) and Rosenthal (1978) emphasize descriptive statistics—such as the mean effect size across a set of studies. This approach is useful not only for summarizing previous research findings but also for disentangling conflicting results and conclusions where the conflict has arisen from some studies showing statistical significance and others failing to do so.

An excellent recent example of this systematic approach to literature review is Hyde's (1981) meta-analysis of previous studies of whether males or females are superior in terms of several dimensions of cognitive ability. Authors of previous qualitative literature reviews had concluded that differences in various abilities were "well-established." However, Hyde found very small effect sizes. Hyde suggested that traditional literature reviews based simply on the number of studies yielding statistically significant results may have misleadingly communicated the impression that the moderately consistent statistically significant sex differences were large when in fact they explained only from 1 to 4% of the variance and averaged less than .5 of the population standard deviation. Hyde concluded that, "of course, a small effect might still be a important one. But at least the reader would have the option of deciding whether a statistically significant effect was large enough to merit further attention, either in teaching or in research" (p. 900).

A marketing meta-analysis that focused on effect size was Clarke's (1976) review of research assessing the duration of advertising effects on sales. Clarke's award-winning meta-analysis made an impact because his prime focus was on three substantive questions: how long do advertising effects last; do other variables interact with those effects; and, if so, how do these interacting variables affect advertising carryover? Clarke analyzed 69 studies, including some for which the effects of advertising were not statistically significant. This meta-analysis yielded several important insights not available from a more traditional qualitative literature review (e.g., Pollay 1979). First, the results indicated that the estimate of the duration of advertising effect was contingent upon the data interval. Shorter intervals (weekly, monthly, or bimonthly) indicated shorter estimates of the duration of advertising effects than longer data intervals (quarterly, annually). Perhaps most important, Clarke was able to conclude that, contrary to past beliefs, advertising effects are likely to last for no more than three to nine months and not years. Clarke summarized by stating that, although he had to make some subjective decisions in

order to produce comparable model specifications, "In isolation, none of the papers gives a satisfactory answer to the question of how long advertising affects sales. By putting them together, as has been done here, one achieves greater confidence in the result" (p. 355).

Several meta-analyses in marketing research have been reported recently. Yu and Cooper (1983) analyzed the effects of several variables on survey response rates after examining 497 response rates from 93 research studies. One conclusion was that, as would be expected intuitively, personal and telephone interviews obtained higher rates of response than mail surveys. However, Yu and Cooper's meta-analysis was able to estimate the size of that and other effects as well as support their presence. Sudman and Bradburn (1974) performed an extensive meta-analysis which investigated the influence of 46 independent variables on response effects. Other recent meta-analyses in marketing research include investigations of 37 multiattribute attitude model studies (Farley, Lehman, and Ryan 1981), four studies examining the Howard-Sheth theory of buyer behavior (Farley, Lehman, and Ryan 1982), 28 studies of price perception (Monroe and Krishnan 1983), and seven studies of the relationship of information search and prior product experience to familiarity (Reilly and Conover 1983).

A systematic meta-analysis can go beyond traditional literature reviews which focus on statistical significance and, in fact, can give a more objective and sometimes different description of results. For example, Rousseau and Redfield's (1980) meta-analysis of the effects of cognitive-level questions on achievement test scores revealed an average effect size of a half of a standard deviation, whereas a traditional analysis of the same literature indicated no effect (Winne 1979). Cooper and Rosenthal (1980) conducted an experiment in which 39 professional researchers analyzed seven studies in either a traditional qualitative manner or with a meta-analysis. The researcher subjects were asked to focus on the average effect size in terms of a Z-score and the average statistical probability of such an effect. Even with this relatively small number of studies to review, the qualitative reviewers formed much different and much less correct impressions about the presence and nature of the relationship between the two variables addressed in the seven studies. Finally, in addition to affording potentially greater objectivity, the use of effect size measures in meta-analysis can suggest point values or ranges that can be compared in subsequent empirical research.

Strong Inference

Although rigorous meta-analyses may increase the likelihood that point value or range predictions can be formulated such that a test of a given theory or hypothesized explanation can go beyond rejections of the null hypothesis, few areas in marketing and consumer research are amenable to such precision at the present time (see Houston, Peter, and Sawyer 1983). However, some situations may at least allow a sorting out of the best currently available theoretical explanation or model from several alternatives.

Platt (1964) advocates strong inference as a useful procedure to augment conventional tests against the null hypothesis. This approach involves comparing competing hypotheses with each other where support for one hypothesis (theory) implies rejection of others. The process of strong inference includes the following steps: (1) devising alternative hypotheses, (2) devising a crucial experiment (or several of them) with alternative possible outcomes each of which will, as nearly as possible, exclude one or more of the hypotheses or explanations, (3) carrying out the experiment so as to get a clean result, (4) recycling the procedure, making subhypotheses or sequential hypotheses to refine the possibilities that remain, and so on. Though the approach sounds simple, much ingenuity clearly is needed to implement this research strategy. However, several examples of the approach are reported in the marketing and consumer behavior literature.

An excellent example of strong inference in hypothesis testing is the investigation of the low-ball technique by Cialdini et al. (1978). The authors observed that automobile sales dealers induce final compliance by getting customers to decide initially to purchase at a lower price and then to retain that compliance when the price advantage is removed. Cialdini et al. used a strong inference design and the results supported an explanation that initial commitment creates a resistance to change in future behavior but not necessarily a more positive attitude. At least as important in terms of strong inference is the fact that the results also rejected the plausibility of the other three explanations of the obtained low-ball effect. Burger and Petty (1981) further refined the conclusions of Cialdini et al. with a strong inference experiment which supported an explanation that an unfulfilled obligation to the person requesting the behavior, not a commitment to the initial target behavior, is responsible for the effectiveness of the low-ball technique.

Another strong inference design directly confronted the Fishbein belief-evaluation multiattribute attitude model with the adequacy-importance approach (Bettman, Capon, and Lutz 1975). This study examined how role-playing subjects formed attitudes toward fictitious brands from given attribute information. The authors used within-individual analyses of variance and ω^2 estimates of explained variance to classify individuals on the basis of how attribute information was utilized. Their research revealed that the multiplicative model was by far the best description of the individuals' information processing and that the Fishbein model was superior to the adequacy-importance model.

Even if use of a strong inference design to test alternative theories is not feasible, one may at least be able to compare a sample result with the value predicted by a given theory or model instead of simply testing whether the result is statistically significantly different from zero. In addition, the predictions of competing or alternative

models can be compared with each other (Armstrong 1979). One such system of statistical analysis is Jöreskog and Sörbom's (1978) maximum likelihood estimation of structural equations to test causal models involving unobservable variables (Bagozzi 1980; Bentler 1980). This approach requires explicit specification of the complex interrelationships among measured and unobservable variables and thus strong theory is needed. Sawyer and Page (1983) summarize how various measures of effect size can augment statistical tests of the fit of sample data to theoretical models.

LIMITATIONS OF OUR RECOMMENDATIONS

We have argued for practices and priorities which differ from current conduct and reporting of empirical research in marketing. Statistics should be used to illuminate rather than obscure data, and we hope that our recommendations can help to achieve this goal. However, we also recognize that there are limitations and problems with any type of hypothesis testing and our recommendations are no exception. In this section we briefly review some of these problems.

We have argued for increased use and reporting of *descriptive statistics* in marketing research. Though such reporting conflicts with the limited space in journals, space constraints should not prevent the inclusion of sufficient information for replication and/or inclusion of the study in a subsequent meta-analysis If journal space constraints preclude the complete description of a study's results, perhaps the journal could require and store pertinent method information, data, and statistics to aid inquiring researchers. We acknowledge, however, that even simple descriptive statistics can sometimes be misleading. For example, averaging many individuals who exhibit "all-at-once" learning patterns, albeit at a varying number of trials, would result in the incorrect conclusion that individuals learn at a gradual rate (Baloff and Becker 1967).

There are several difficulties in the quantification, interpretation, and generalization of effect size measures. Some such measures estimate the ratio of explained to total variance. In quantifying the amount of explained variance (such as R^2 or ω^2), researchers must realize that total variance is increased by measurement and treatment unreliability, heterogeneous subjects, and poorly controlled research procedures (O'Grady 1982; Sechrest and Yeaton 1981a,b). Experimental researchers also can influence the amount of explained variance by restricting or magnifying the manipulation of an independent variable. Independent variables which are qualitative or categorical present particular interpretation problems. Such variables often have no conceptually meaningful or practically important characteristics in common within or across studies; the number of "levels" of such variables is infinite and any estimates of the "size" of their effects are very difficult to interpret. Finally, the problems of the influence of individual characteristics of particular studies and manipulations within a study make it very difficult to generalize effect sizes meaningfully or to compare them across a set of different studies (such as in a meta-analysis). However effect sizes are estimated, these descriptive statistics are more generalizable if the levels of the independent variable are a random subset of all levels of interest (Glass and Hakstian 1969) and orthogonal to other independent variables (Green, Carroll, and DeSarbo 1978; LaTour 1981a).

Fortunately, other measures of effect size are available. Rosenthal (1978) discusses the advantages and disadvantages of nine relatively simple methods of summarizing results including three estimates of effect size. These methods include adding *t*-test statistics, *Z*-values, and weighted *Z*-values. LaTour (1981a,b) recommends the use of a contrast estimate to quantify effect size because it eliminates many of the problems of explained variance estimates. Glass, McGaw, and Smith (1981, p. 102) recently concluded that, "The findings of comparative experiments are probably best expressed as standardized mean differences between pairs of treatment differences." Most of these methods that do not estimate the proportion of explained variance seem most appropriate for simple research designs and are difficult to use and interpret with more complex designs (Glass and Hakstian 1969).

Some limitations of our other recommendations should also be noted. Though we believe that *replication research* is very important, recognition for conducting replications seems to be lacking in marketing research. Also, it is very unlikely that all sources of variance in research involving human subjects can be specified or controlled. Thus, replications can never exactly duplicate prior research conditions, and different results may be obtained. Such conflicting results can lead to confusion rather than consensus. Of course, confusion is better than the acceptance of a single result as conclusive, and subsequent meta-analyses may be able to determine the source of the conflict in results.

We believe *Bayesian hypothesis testing* is useful, but also recognize that researchers need to have a high degree of mathematical sophistication to understand and apply the approach. It is clearly not an approach which is amenable to canned computer programs and is thus difficult for researchers to use.

In addition to the problem of meaningfully comparing effect sizes, a *meta-analysis* often encounters other formidable obstacles. One problem is the search for a census of studies including the unpublished ones that are likely to have smaller effect sizes. For studies that are available, information is often insufficient for calculating effect sizes and study authors must be contacted. Unfortunately, it is also often difficult to obtain sufficiently detailed descriptions of the study method and to code these study characteristics so that their effects can be assessed in the meta-analysis. Small samples and confounded study characteristics make it difficult to disentangle main effects across studies, as well as complex

interactions. An opposite problem is that, if all surveyed studies use the same procedure, the effect of that method cannot be assessed (e.g., Cartwright 1973). One important outcome of a meta-analysis might be a specification of types of studies that would fill a void and allow an examination of the effects of variables that cannot currently be meaningfully evaluated.

It should be obvious that a meta-analysis, though quantitative, depends on many subjective researcher decisions and affords much opportunity for disagreement. Perhaps because the publication of a meta-analysis carries an aura of finality, researchers very commonly disagree about the many decisions involved in a meta-analysis and, hence, challenge the conclusions. For example, Stanley and Benbow (1982) challenged Hyde's (1981) meta-analysis of gender differences in quantitative ability and Weinberg and Weiss (1982) disputed some of the analysis decisions in Clarke's (1976) meta-analysis of advertising carryover as well as the statistical validity of his conclusions.[4]

Finally, though *strong inference designs* are superior to test against the null hypothesis, often theories are incommensurable and hence cannot be confronted empirically. In addition, even strong inference designs can obtain conflicting results. For example, Mazis, Ahtola, and Klippel (1975) compared four formulations of multiattribute attitude models and concluded that the adequacy-importance model yielded better predictions than the Fishbein model. This conclusion conflicts with the findings of Bettman, Capon, and Lutz (1975).

Though the preceding discussion is by no means a complete list of limitations, the problems noted should serve as a reminder of one critical fact: *there is no universal approach to hypothesis testing which can guarantee a meaningful empirical test or offer fully objective analysis and description of results*. Some approaches are better than others for particular problems. As we have illustrated, biases in choosing an approach and the decisions made in implementing it have an extremely important influence on conclusions from the data. Thus, if possible, researchers ought to use multiple approaches to testing hypotheses and reporting the results.

SUMMARY

Several issues related to the interpretation and value of statistical significance testing are reviewed. Although properly applied statistical significance tests are useful aids in drawing inferences and for signalling relationships which need further study, they are not sufficient for falsifying hypotheses or judging research results. Despite the fact that many of these ideas have been discussed previously, many researchers, including those in marketing, continue to ignore them. Attention should be placed on the data themselves and their descriptions. Instead of relying solely on classical inferential statistics, researchers should make added use of replication, Bayesian statistics, meta-analysis, and strong inference to provide more meaningful examination of theoretical questions in marketing research.

[4]Though the statistical models involved in the exchange between Weinberg and Weiss and Clarke (1982) are very sophisticated, the arguments pertain to important basic ideas discussed in this article about statistical power, whether failure to reject the null hypothesis implies that the null hypothesis is true, and the need for testing results against theoretically based point value predictions instead of merely comparing results against a zero point null hypothesis.

REFERENCES

Armstrong, J Scott (1979), "Advocacy and Objectivity in Science," *Management Science*, 25 (May), 423–38.

Assael, Henry and John Keon (1982), "Nonsampling vs. Sampling Errors in Survey Research," *Journal of Marketing*, 46 (Spring), 114–23.

Bagozzi, Richard P. (1978), "Salesforce Performance and Satisfaction as a Function of Individual Difference, Interpersonal, and Situational Factors," *Journal of Marketing Research*, 15 (November), 517–31.

——— (1980), *Causal Models in Marketing*. New York: John Wiley & Sons, Inc.

Bakan, David (1966), "The Test of Significance in Psychological Research," *Psychological Bulletin*, 66 (December), 423–37.

Baloff, Nicholas and Selwyn Becker (1967), "On the Futility of Aggregating Individual Learning Curves," *Psychological Reports*, 20, 183–91.

Banks, Seymour (1965), *Experimentation in Marketing*. New York: McGraw-Hill Book Company.

Bentler, P. M. (1980), "Multivariate Analysis with Latent Variables: Causal Modeling," in *Annual Review of Psychology*, Vol. 31, M. R. Rosenzweig and L. W. Porter, eds. Palo Alto, CA: Annual Reviews.

Bettman, James R. (1974), "Toward a Statistics for Consumer Decision Net Models," *Journal of Consumer Research*, 1 (June), 71–80.

———, Noel Capon, and Richard J. Lutz (1975), "Cognitive Algebra in Multi-Attribute Attitude Models," *Journal of Marketing Research*, 12 (May), 151–64.

Blattberg, Robert C. (1979), "The Design of Advertising Experiments Using Statistical Decision Theory," *Journal of Marketing Research*, 16 (May), 191–202.

Brown, Rex V. (1967), "Evaluation of Total Survey Error," *Journal of Marketing Research*, 4 (May), 117–27.

Brown, Stephen W. and Corbett F. Gaulden, Jr. (1980), "Replication and Theory Development," in *Theoretical Developments in Marketing*, C. W. Lamb, Jr. and P. M. Dunne, eds. Chicago: American Marketing Association, 240–3.

Burger, Jerry M. and Richard E. Petty (1981), "The Low-Ball Compliance Technique: Task or Person Commitment?," *Journal of Personality and Social Psychology*, 40 (March), 492–500.

Carlsmith, J. Merrill, Phoebe C. Ellsworth, and Elliot Aronson (1976), *Methods of Research in Social Psychology*. Reading, MA: Addison-Wesley.

Cartwright, Dorwin (1973), "Determinants of Scientific Progress: The Case of Research on the Risky Shift," *American Psychologist*, 28 (March), 222–31.

Carver, Ronald P. (1978), "The Case Against Statistical Significance Testing," *Harvard Educational Review*, 48 (August), 278–399

Cialdini, Robert B., John T. Cacioppo, Rodney Bassett, and John A. Miller (1978), "Low-Ball Procedure for Producing Compliance Commitment then Cost," *Journal of Personality and Social Psychology,* 36 (May), 463–76.

Clarke, Darral G (1976), "Econometric Measurement of the Duration of Advertising Effect on Sales," *Journal of Marketing Research,* 13 (November), 345–57.

——— (1982), "A Reply to Weinberg and Weiss," *Journal of Marketing Research,* 19 (November), 592–4.

Cohen, Jacob (1977), *Statistical Power Analysis for the Behavioral Sciences.* New York. Academic Press.

Cook, Thomas D and Donald T. Campbell (1979), *Quasi-Experimentation: Design and Analysis for Field Settings.* Chicago: Rand-McNally.

Cooper, Harris M. and Robert Rosenthal (1980), "Statistical Versus Traditional Procedures for Summarizing Research Findings," *Psychological Bulletin,* 87 (May), 442–9.

Cronbach, Lee J and R. E. Snow (1977), *Aptitudes and Instructional Methods. A Handbook for Research on Interactions.* New York: Irvington.

Della Bitta, Albert J , Kent B. Monroe, and John M. McGinnis (1981), "Consumer Perceptions of Comparative Price Advertisements," *Journal of Marketing Research,* 18 (November), 416–27.

Dodson, Joe A., Alice M. Tybout, and Brian Sternthal (1978), "Impact of Deals and Deal Retraction on Brand Switching," *Journal of Marketing Research,* 15 (February), 72–81.

Edwards, Ward, Harold Lindman, and Leonard J Savage (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review,* 70 (May), 193–242.

Eskin, Gerald J. and Penny H. Baron (1977), "Effects of Price and Advertising in Test-Market Experiments," *Journal of Marketing Research,* 14 (November), 499–508

Farley, John U., Donald R. Lehman, and Michael J Ryan (1981), "Generalizing from 'Imperfect' Replication," *Journal of Business,* 54 (October), 597–610.

———, ———, and ——— (1982), "Patterns in Parameters of Buyer Behavior Models: Generalizing from Sparse Replication," *Marketing Science,* 1 (Spring), 181–204.

Ginter, James, Martha Cooper, Carl Obermiller, and Thomas Page (1981), "The Design of Advertising Experiments: An Extension," *Journal of Marketing Research,* 18 (February), 120–3.

Glass, Gene V. and A. Ralph Hakstian (1969), "Measures of Association in Comparative Experiments: Their Development and Interpretation," *American Educational Research Journal,* 6 (May), 403–14.

———, Barry McGaw, and Mary Lee Smith (1981), *Meta-Analysis in Social Research.* Beverly Hills, CA: Sage Publications.

Green, Paul E., J. Douglas Carroll, and Wayne S DeSarbo (1978), "A New Measure of Predictor Variable Importance in Multiple Regression," *Journal of Marketing Research,* 15 (August), 356–60.

Greenwald, Anthony G. (1975a), "Consequences of Prejudice Against the Null Hypothesis," *Psychological Bulletin,* 82 (January), 1–19.

——— (1975b), "Does the Good Samaritan Parable Increase Helping? A Comment on Darley and Batson's No-Effect Conclusion," *Journal of Personality and Social Psychology,* 32 (October), 578–83

——— (1975c), "Significance, Nonsignificance, and Interpretation of an ESP Experiment," *Journal of Experimental Social Psychology,* 11 (March), 180–91

Henkel, Ramon E. (1976), *Tests of Significance* Beverly Hills, CA: Sage Publications.

Houston, Michael J., J. Paul Peter, and Alan G. Sawyer (1983), "The Role of Meta-Analysis in Consumer Behavior Research," in *Advances in Consumer Research,* Vol. 10, R P. Bagozzi and A. M. Tybout, eds. Ann Arbor, MI Association for Consumer Research.

Hyde, Janet Shibley (1981), "How Large Are Cognitive Gender Differences?: A Meta-Analysis Using ω^2 and d," *American Psychologist,* 36 (August), 892–901.

Iverson, Gudmund R. (1970), "Statistics According to Bayes," in *Sociological Methodology,* Edgar F. Borgatta and George W. Bohrnstedt, eds San Francisco: Jossey-Bass, 185–99.

Jöreskog, Karl G. and Dag Sörbom (1978), *LISREL · Analysis of Linear Structural Relationships by the Method of Maximum Likelihood,* Version IV, Release 2 Chicago· National Educational Resources, Inc.

Kish, Leslie (1959), "Some Statistical Problems in Research Design," *American Sociological Review,* 24 (June), 328–38.

Krugman, Herbert E (1971), "Brain Wave Measures of Media Involvement," *Journal of Advertising Research,* 11 (February), 3–9

LaTour, Stephen A. (1981a), "Effect Size Estimation: A Commentary on Wolf and Bassler," *Decision Sciences* (January), 136–41.

——— (1981b), "Variance Explained: It Measures Neither Importance nor Effect Size," *Decision Sciences* (January), 150–60.

Layton, R. D. and B. Turnbull (1975), "Belief, Evaluation, and Performance on an ESP Task," *Journal of Experimental Social Psychology,* 11 (March), 166–79

Leone, Robert P. and Randall L. Schultz (1980), "A Study of Marketing Generalizations," *Journal of Marketing,* 44 (Winter), 10–18

Levitt, Theodore (1972), "Industrial Purchasing Behavior A Bayesian Reanalysis," *Journal of Business Administration,* 4 (Fall), 79–81

Lipstein, Benjamin (1975), "In Defense of Small Samples," *Journal of Advertising Research,* 15 (February), 33–40

Lykken, David T. (1968), "Statistical Significance in Psychological Research," *Psychological Bulletin,* 70 (September), 151–9.

Mayer, Charles (1970), "Assessing the Accuracy of Marketing Research," *Journal of Marketing Research,* 7 (August), 285–91.

Mazis, Michael, Olli T Ahtola, and R Eugene Klippel (1975), "A Comparison of Four Multi-Attribute Models in the Prediction of Consumer Attitudes," *Journal of Consumer Research,* 2 (June), 38–52.

McGuire, William J. (1973), "The Yin and Yang of Progress in Social Psychology: Seven Koan," *Journal of Personality and Social Psychology,* 26 (June), 446–56.

Meehl, Paul E. (1967), "Theory Testing in Psychology and Physics A Methodological Paradox," *Philosophy of Science,* 16 (June), 103–15

——— (1978), "Theoretical Risks and Tabular Asterisks. Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology," *Journal of Consulting and Clinical Psychology,* 46, 806–84

Meyer, Donald L (1974), "Statistical Tests and Surveys of Power: A Critique," *American Educational Research Journal,* 11 (Spring), 179–88.

Monroe, Kent B and R Krishnan (1983), "The Effect of Price

on Subjective Product Evaluations: A Synthesis of Outcomes," in *Advances in Consumer Research,* Vol. 10, R. P. Bagozzi and A. M. Tybout, eds. Ann Arbor, MI: Association for Consumer Research.

O'Grady, Kevin E. (1982), "Measures of Explained Variance: Cautions and Limitations," *Psychological Bulletin,* 92 (November), 766–77.

Peter, J. Paul (1983), "Some Philosophical and Methodological Issues in Consumer Research," in *Marketing Theory: The Philosophy of Marketing Science,* Shelby D. Hunt, ed. Homewood, IL: Richard D. Irwin.

Phillips, Lawerence D. (1973), *Bayesian Statistics for Social Sciences.* London: Thomas Nelson.

Platt, John R (1964), "Strong Inference," *Science,* 146 (October 16), 347–53.

Pollay, Richard W. (1979), "Lydiametrics: Applications of Econometrics to the History of Advertising," *Journal of Advertising History,* 1, 3–18.

Reilly, Michael D. and Jerry N. Conover (1983), "Meta-Analysis: Integrating Results from Consumer Research Studies," in *Advances in Consumer Research,* Vol. 10, R P Bagozzi and A M. Tybout, eds. Ann Arbor, MI: Association for Consumer Research

Roberts, Harry V. (1963), "Bayesian Statistics in Marketing," *Journal of Marketing,* 27 (January), 1–4.

Rosenthal, Robert (1969), "Interpersonal Expectations: Effects of the Experimenter's Hypothesis," in *Artifact in Behavioral Research,* Robert Rosenthal and Ralph L Rosnow, eds. New York: Academic Press, 181–277.

——— (1978), "Combining Results of Independent Studies," *Psychological Bulletin,* 85 (December), 185–93.

——— (1979), "The 'File Drawer Problem' and Tolerance for Null Results," *Psychological Bulletin,* 86 (March), 638–41

——— and John Gaito (1963), "The Interpretation of Levels of Significance by Psychological Researchers," *Journal of Psychology,* 55, 33–8.

——— and Donald B Rubin (1982), "Comparing Effect Sizes of Independent Studies," *Psychological Bulletin,* 92 (September), 500–4.

Rousseau, E W. and D L Redfield (1980), "Teacher Questioning," *Evaluation in Education, An International Review Series.* 4, 51–2.

Ryans, Adrian B. and Charles B. Weinberg (1979), "Territory Sales Reponse," *Journal of Marketing Research,* 16 (November), 453–65.

Sawyer, Alan G. and A. Dwayne Ball (1981), "Statistical Power and Effect Size in Marketing Research," *Journal of Marketing Research,* 18 (August), 275–90

——— and Thomas J. Page, Jr. (1983), "Incremental Goodness of Fit Indices in Structural Equation Models in Marketing Research," paper presented at the AMA special Conference on Causal Modeling, Sarasota, FL, March 2.

Schlaifer, Robert (1961), *Introduction to Statistics for Business Decisions.* New York: McGraw-Hill Book Company.

Sechrest, Lee and William Yeaton (1981a), "Empirical Bases for Estimating Effect Size," in *Reanalyzing Program Evaluations: Policies and Practices,* R. F. Boruch, P M. Wortman, and D. S. Cordray, eds. Ann Arbor: University of Michigan Institute for Social Research

——— and ——— (1981b), "Estimating Magnitudes of Experimental Effects," unpublished manuscript, University of Michigan Institute for Social Research, Ann Arbor.

Selvin, Hanan C. (1957), "A Critique of Tests of Significance in Survey Research," *American Sociological Review,* 22 (October), 519–27

Smith, N. C , Jr. (1970), "Replication Studies: A Neglected Aspect of Psychological Research," *American Psychologist,* 25 (October), 970–5

Stanley, Julian C. and Camilla P. Benbow (1982), "Huge Sex Ratios at Upper End," *American Psychologist,* 37 (August), 972

Stevens, S. S (1971), "Issues in Psychophysical Measurement," *Psychological Review,* 78 (September), 426–50.

Sudman, Seymour and Norman M. Bradburn (1974), *Response Effects in Surveys: A Review and Synthesis.* Chicago: Aldine.

Tversky, Amos and Daniel Kahneman (1971), "Belief in the Law of Small Numbers," *Psychological Bulletin,* 76 (August), 105–10.

Weinberg, Charles B. and Doyle L. Weiss (1982), "On the Econometric Measurement of the Duration of Advertising Effects on Sales," *Journal of Marketing Research,* 19 (November), 585–91.

Winch, Robert F. and Donald T. Campbell (1969), "Proof? No. Evidence? Yes The Significance of Tests of Significance," *American Sociologist,* 4 (May), 140–3.

Winne, P. H. (1979), "Experiments Relating Teacher's Use of Higher Cognitive Questions to Student Achievement," *Review of Educational Research,* 49, 13–50.

Yu, Julie and Harris Cooper (1983), "A Quantitative Review of Research Design Effects in Response Rates to Questionnaires," *Journal of Marketing Research,* 20 (February), 36–44

Zeisel, Hans (1955), "The Significance of Insignificant Differences," *Public Opinion Quarterly,* 17 (Fall), 319–21.