A Critique of the Usefulness of Inferential Statistics in Applied Behavior Analysis

B. L. Hopkins, Brian L. Cole, and Tina L. Mason
Auburn University

Researchers continue to recommend that applied behavior analysts use inferential statistics in making decisions about effects of independent variables on dependent variables. In many other approaches to behavioral science, inferential statistics are the primary means for deciding the importance of effects. Several possible uses of inferential statistics are considered. Rather than being an objective means for making decisions about effects, as is often claimed, inferential statistics are shown to be subjective. It is argued that the use of inferential statistics adds nothing to the complex and admittedly subjective nonstatistical methods that are often employed in applied behavior analysis. Attacks on inferential statistics that are being made, perhaps with increasing frequency, by those who are not behavior analysts, are discussed. These attackers are calling for banning the use of inferential statistics in research publications and commonly recommend that behavioral scientists should switch to using statistics aimed at interval estimation or the method of confidence intervals. Interval estimation is shown to be contrary to the fundamental assumption of behavior analysis that only individuals behave. It is recommended that authors who wish to publish the results of inferential statistics be asked to justify them as a means for helping us to identify any ways in which they may be useful.

Key words: research methods, data analysis, inferential statistics

Several writers (Baer, 1977; Johnston & Pennypacker, 1993; Michael, 1974; Sidman, 1960) have cautioned against the use of inferential statistics in behavior analysis research, but many other authors have recommended that behavior analysts use inferential statistics in making decisions about possible experimental effects (Edgington, 1980, 1982; Gentile, Roden, & Klein, 1972; Gottman, 1981; Gottman & Glass, 1978; Hartmann, 1974; Hartmann et al., 1980; Horne, Yang, & Ware, 1982; Huitema, 1986; Jones, Vaught, & Weinrott, 1977; Kazdin, 1976; Keselman & Leventhal, 1974; Kratochwill, 1974; Kratochwill et al., 1974; Kratochwill & Levin, 1980; Mainstone &

Levi, 1987; Notz, Boschman, & Tax, 1987; Pfadt, Cohen, Sudhalter, Romanczyk, & Wheeler, 1992; Pfadt & Wheeler, 1995; Thoresen & Elashoff, 1974).

The encouragement for using inferential statistics in behavior analysis is consistent with behavioral science's broad adoption of Fisher's (1925) prescription for making decisions about differences in sets of data. Gigerenzer and Murray (1987) concisely characterize Fisher's position: "Scientific knowledge comes only from inductive inference. . . . Inductive inference is chiefly disproving null hypotheses. . . . Therefore, all scientists must try to disprove null hypotheses" (p. 11). Fisher's writings emphasized using inferential statistics as a method of disproving null hypotheses: "we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result" (p. 14). According to this view, an independent variable in an experiment is typically judged to have an effect on a dependent variable if changes or differences in the dependent vari-

The initial draft of this manuscript was written while the first author was on a sabbatical hosted by the Department of Psychology at Emory University. We appreciate wise comments and suggestions of David G. Born, Dudley F. McGlynn, and Barry S. Parsonson. Several reviewers of previous drafts of this manuscript also made many useful comments.

Correspondence concerning this paper should be addressed to B. L. Hopkins, Department of Psychology, Auburn University, Auburn, Alabama 36849.

able are found to be statistically significant.

The influence of Fisher's (1925) view has been broad and deep. Behavioral scientists commonly use some form of inferential statistics to decide whether effects of independent variables or the relationships between variables are important. For example, we examined all of the research articles published in selected issues of the same year in such diverse journals as *Child Development* (Somerville, 1994), *Journal of Applied Psychology* (Schmitt, 1994), *Journal of Clinical and Experimental Neuropsychology* (Costa & Rourke, 1994), *Journal of Educational Psychology* (Levin, 1994), *Journal of Experimental Psychology: Animal Behavior Processes* (Hulse, 1994), *Journal of Experimental Psychology: General* (Hunt, 1994), *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Rayner, 1994), and *Journal of Social Psychology* (Doob, 1994). Of the 102 articles, 98 employed inferential statistics in deciding whether sets of data were different from each other or from some hypothesized value or were importantly related.

All of the four articles that were exceptions were special cases. One article in the *Journal of Educational Psychology* was a descriptive study of college students' explanations for why they took notes in classes (Van Meter, Yokoi, & Pressley, 1994). Two research papers published in the *Journal of Social Psychology* may not have used inferential statistics; one of these reported simple correlations (Williams, 1994) and the other reported factor loadings on a questionnaire (Dugbarney, 1994). Inferential statistics could have been employed in the research described in these social psychology papers, but such use was not mentioned. Last, a case study reported in the *Journal of Clinical and Experimental Neuropsychology* simply described a procedure and did not use inferential statistics (Wolff, Sass, & Keidan, 1994).

Some of the papers published in the major behavior analysis research journals, the *Journal of the Experimental Analysis of Behavior* and the *Journal of Applied Behavior Analysis*, report the use of inferential statistics. In Volume 40 of the *Journal of the Experimental Analysis of Behavior* (Nevin, 1983), 10 of 25 research articles employed inferential statistics. In Volume 60 of the same journal (Branch, 1993), 12 of 24 research reports used inferential statistics. In the second and third issues of Volume 16 of the *Journal of Applied Behavior Analysis* (Barlow, 1983), six of 15 research reports used inferential statistics, and in the second and third issues of Volume 26 (Neef, 1993), only one article of 23 used inferential statistics.

Many practices of our institutions that are involved in research support the use of inferential statistics. For example, federal granting agency study sections are groups of scientists who make recommendations to fund or not to fund research proposals. The first author has sat on study sections of four different federal granting agencies. All the study sections included at least one and frequently several reviewers who particularly examined the adequacy of proposed statistical methods for making decisions about the importance of effects and relationships that might result from the research if it were carried out. This author kept no descriptive statistics on several hundred funding recommendations made by the study sections. However, he recalls no instance in which experimental research was recommended for funding if it did not propose to use inferential statistics in making decisions about effects. In addition, a number of recommendations against funding were based primarily on proposals' failing to describe the planned use of the inferential statistics judged to be the proper ones by study section members.

Our institutions of higher education broadly support the use of inferential statistics. Undergraduate training programs in the behavioral sciences com-

monly require that students take one course in statistics with the advanced course topics dealing with inferential statistics. Graduate training programs often require two or more courses with major emphases on inferential statistics.

What are inferential statistics supposed to accomplish for behavioral scientists? Perhaps the use of inferential statistics is so common that their supposed importance is easily taken for granted; many statistics texts (e.g., Howell, 1992; Spence, Cotton, Underwood, & Duncan, 1990) introduce the methods of basing decisions about data on inferential statistics without explaining why that is a recommended practice. Other authors imply that such use of statistics provides a relatively good or objective means for making decisions about effects and relationships. For example, Hurlburt (1994) states, "Statistics are the best tools available for deciding whether inductive statements should be considered 'true' " (p. 3). Hays (1963) stated, "Faced . . . with an experiment he (sic) can conduct only once . . . armed with this [statistical] information, the experimenter is in a better position to decide, if he must, what he will say about the true situation" (p. 11). Jones et al. (1977) extended the arguments for objectivity to applied behavioral research, arguing that inferential statistics afford a degree of objectivity that may be lacking in nonstatistical behavior analysis methods and may overcome lack of agreement between scientists who use nonstatistical approaches.

The thesis of the present paper is that the use of inferential statistics is not an objective tool for examining effects of independent variables and that such statistics add little or nothing beyond what is provided by common behavior analysis methods. It will be further argued that inferential statistics, although inherently harmless, would probably result in poor applied behavioral research.

First, we will critically examine several different ways in which inferential

statistics could be used in applied behavioral research. For each possible usage of inferential statistics, we will discuss certain limitations inherent in them and will strategically contrast the usage to alternative methods of behavior analysis.

Statistical Significance As the Major Consideration

Inferential statistics appear to be most often used in behavioral science research as the primary basis for deciding whether or not some independent variable had an important effect on some dependent variable or whether two variables were importantly related (Atkinson, Furlong, & Wampold, 1982). At the extreme, a research article may report the results of statistical calculations, typically the numerical value of the statistic and the probability of that value, without presenting any data or even numerical summaries of data. Making the result of the statistical test the major basis for decisions about the importance of effects has a flaw that can be easily seen by examining the ways in which statistics are calculated.

The following example and research design are not chosen because they are necessarily ideal or typical but because they conveniently reveal the mechanics of a particular statistical calculation. Suppose a researcher is testing the effectiveness of a training program intended to help mothers of small children influence those children to smile more and cry less. The researcher takes data on how much of the time some children smile, then trains their mothers, and then takes posttraining data to see if the smiling has increased. The researcher wishes to examine whether changes in the data are statistically significant following the treatment. Assume, for the moment, that the data satisfy assumptions sometimes required before statistical tests are conducted.

A *t* test for dependent samples might be used to decide whether the mean of

the differences in the pre- and post-training data are significantly different from zero. The t statistic is calculated from the following equation:

$$t = \frac{\bar{D}}{\frac{SD_D}{\sqrt{n}}},$$

with $n - 1$ degrees of freedom, where $\bar{D}$ is the mean of the differences, SD_D is the standard deviation of the differences, and n is the number of differences. In this case, the number of differences is the same as the number of subjects on whom data were collected. The equation is solved for t , and the result is compared to a table that gives the values of t necessary to achieve statistical significance at certain levels of confidence, with the value of t necessary to obtain statistical significance decreasing with increasing n . The larger the t value resulting from the use of this equation, the better the chances for achieving a chosen level of statistical significance.

Examining the equation, t is seen to be a function of the magnitude of the mean of the differences, the standard deviation or variability of the differences, and the sample size. It is important to note, then, that a statistically significant t can be achieved in several different ways. First, a large effect (i.e., a large $\bar{D}$) with some variability and a modest sample size might be found to be statistically significant. Second, a smaller effect, if it were accompanied by less variability (i.e., a smaller SD), might also be found to be statistically significant. Finally, a small effect, with considerable variability, might be found to be statistically significant if the sample size is large (i.e., n is large). In fact, the equation is such that t will increase as the sample size increases given any constant ratio of mean difference to variability. In this last way statistical significance could be achieved, it should be noted that not only does the calculated value of t increase with increasing n , but the tabled values of t required for a given level of signifi-

cance also decrease with increasing values of n . In other words, if a t is not statistically significant given a particular effect size and variability, it can be made significant with no increase in the effect size or decrease in the variability simply by increasing the n .

To make this fact concrete, consider some numerical values for the different terms in the equation. Suppose the mean difference is .40, the training program produces small effects or effects that are so variable over children that the mean effect is small, the standard deviation of the differences is 1.0, and the number of children in the experiment is 9. If these numbers are inserted into the equation and it is solved for t , the result is 1.2. If this result is compared to a table of the t distribution, we find that we should expect to get a t value this large, even if the training has absolutely no effect, more than 20 times in 100 simply by "chance" or the "random" variations that are occurring in the data. However, if we have the resources to simply include more mothers and children in our training and get similar changes in the data on the children's smiling following training, our statistical result changes. Without an increase in the size of the mean of the changes or any decrease in the variability, but with data on say 36 children, we obtain 2.4 for the value of t . That value is now more than enough to achieve statistical significance.

From one perspective, this relationship among the four terms of the equation seems reasonable. In applied research we want to identify variables that have effect sizes that are large in relation to variability. We want the effects to have subject generality in the multiple-subject case and be durable over time in the individual-subject multiple-observation case.

From a different perspective, the relationship among the four terms of the equation is seriously problematic. Given a certain variability, even quite small differences will be found to be statistically significant as the sample

size becomes large. If statistical significance is used as a major determinant in deciding the importance of effects, there is clearly the possibility that an independent variable will be declared important, not because it has a large effect on the dependent variable in relation to the size of the variability, nor even because some small effect is obtained for a large proportion of the subjects, but simply because data from a large number of subjects are being considered. This is a mechanism by which statistical significance, if employed as a criterion, could lead to quite small effects being considered important (Baer, 1977).

It is important to note that the relationship between sample size and statistical significance holds for other statistics. It holds for the applications of analysis of variance, for chi-square tests, and for the various nonparametric statistical inference tests based on the binomial distribution. The numerical values of all of these inferential statistics are affected by sample sizes in substantially the same ways as is the value of t .

This is not a new observation. Hays (1963) noted, "Virtually any study can be made to show significance if one uses enough subjects, regardless of how nonsensical the content may be. There is surely nothing on earth that is completely independent of anything else" (p. 340). Similarly, Hurlburt (1994) wrote, "Because the null hypothesis is seldom (if ever) actually exactly true, a large enough sample size will almost always produce significant results" (p. 214).

We have heard some researchers argue that not all independent variables have even small effects on all dependent variables. Disagreement about whether all independent variables have some effect on all dependent variables is, of course, practically unimportant and beyond convenient arbitration. The important point is that statistical significance is a function of sample size and that even very small effects will

be found to be significant if sufficiently large sample sizes are used.

If a researcher or journal editor or consumer of research tries to decide whether an independent variable has an important effect on a dependent variable, when the decision is based on inferential statistics, the conclusion will vary with sample size. Statistical significance resulting from use of a large sample might occur with small effects that held for only a slender majority of subjects, whereas a lack of significance could result from very large effects that held for all of a few subjects or for only some of many subjects. No objective rules exist for adjusting decisions according to sample sizes. Therefore, the objective result of a test of statistical significance must be subjectively interpreted in relationship to sample size. Even if the calculation of statistical significance is objective, further necessary considerations that would take into account sample sizes are purely subjective.

Some possible consequences of the practice of basing decisions about effects of independent variables primarily on results of statistical tests are obvious and disturbing. Perhaps most important, a large number of independent variables that have relatively small effects will become categorized as being important. At least one possible result of this faulty categorization is that an applied science could become burdened by large numbers of independent variables that are known to have some effect on important behaviors but in fact have only small effects. Such a science would become unable to affect those behaviors in ways that are socially significant. If the standards for research become such that any effect is valued, then the contingencies for researchers seem to encourage identifying more treatments that have any statistically significant effect on behaviors rather than on changing those behaviors in important ways. There will surely be many easily discovered treatments that will have statistically sig-

nificant effects, but perhaps few that change behavior in important ways.

Given that granting agencies typically base funding at least partly on the use of inferential statistics, and that grant funding of research probably allows inclusion of more subjects, there seems to be a real possibility that such funding would contribute to the identification of relatively weaker independent variables as being important. In short, grant funding practices could favor declaring weak treatments to be important.

Behavioral science journals contain many studies of independent variables that are commonly considered important when, in fact, they are trivial. A likely result would be our eventually discounting large segments of research literature and sometimes discarding potentially useful treatments along with those of little value.

Statistical Significance Modified by Other Considerations

Perhaps researchers could first use inferential statistics to decide whether changes in data were statistically significant and then use our more customary methods of behavior analysis to further consider whether the statistically significant changes were practically important. Johnston and Pennypacker (1993), Cooper, Heron, and Heward (1987), and Parsonson and Baer (1986) have described complex nonstatistical methods for evaluating behavior analysis research. A thorough treatment of these methods is beyond the scope of the present paper. However, the methods include consideration of the following: the magnitude of the effect; the variability of the data, usually of a single subject, both before and after some experimental manipulation; the adequacy of the experimental design for the problem being addressed; the value of concluding that an effect is important if, in truth, it is; the value of concluding that an effect is unimportant when, in fact, it is unimportant; the costs of concluding that an effect is im-

portant when, in truth, it is not; the costs of concluding that an effect is not important when, in truth, it is; the value or social significance of the changes in data that occur with the experimental manipulation; the durability of the changes in data; and the number and kinds of subjects for whom socially important changes occur. For much applied behavioral research, we would add that there are other important considerations such as how rapidly data change with the manipulation of experimental conditions and the value of the changes in data in relation to their costs. These considerations are subjective, but their subjectivity is widely known and scrutinized. For convenience we will refer to these as *behavior analysis methods*.

A two-step process of first deciding statistical significance and subsequently deciding practical importance would probably sometimes yield satisfactory decisions about the importance of effects of independent variables. However, it is likely that the more complex behavior analysis methods would sometimes yield decisions that are contrary to the statistical decisions because of the contributions of sample size to statistical significance. Small effects that were statistically significant would be found to be practically unimportant, and statistically nonsignificant effects, perhaps based on a small number of observations or surrounded by considerable variability, would be found to be important. For example, an independent variable might have a powerful and practically important effect on the behavior of some subjects but not others. The value for some but not all subjects would be important. The aggregated effect would yield a nonsignificant statistical value.

If the two steps of the process were reversed so that the complex methods of behavior analysis were used first, would the use of inferential statistics add anything valuable? We think not. If, after weighing all of the considerations common in behavior analysis research, an effect were judged to be

practically important, a further test of statistical significance would be irrelevant. If an experimental manipulation were accompanied by socially important changes in a dependent variable and these changes occurred for a practically important proportion of all subjects and the values of the changes were greater than their costs, finding that the changes were statistically significant seems to add nothing to judgments about the changes. Neither would finding that the changes in the dependent variable were not statistically significant obviate the prior judgments based on the methods of behavior analysis.

The Small-Effect Case

Perhaps even if inferential statistics are not useful in the general case in which it is important to decide if an independent variable has an effect on behavior, they may be useful in identifying small but important effects that are not easily detected by behavior analysis methods. Suppose some independent variable has a small but real effect on the dependent variable, and it is practically important that we not declare this independent variable to be ineffective. Perhaps an inexpensive drug with no known side effects reduces by a small percentage the incidence of strokes among older people. It may be very important that we not declare the drug useless on the basis of our being unable to detect, by behavior analysis methods, any difference in the distributions of strokes by the people taking the drug and those not taking it. The argument is simply that statistical inference may be less likely than our complex behavior analysis methods to reject as unimportant an independent variable that has a small effect that is, indeed, important.

Recall the above argument, that the use of inferential statistics is more likely than complex behavior analysis methods to label unimportant effects as important. This is a different way of saying that inferential statistics are rel-

atively generous in finding that small effects are important. If that is correct, then inferential statistics might be more efficient than the behavior analysis methods for identifying small effects that are important if the weighting given the various behavioral methods is not adjusted for the importance of the problem, the importance of not deciding that the treatment has no effect if it has some effect, the benefits and costs of the effects, and so forth. However, prudent behavior analysis research should adjust the emphasis put on various considerations depending on the nature of the problem being addressed. A very simple example would be reducing the magnitudes of effects that we would declare important if we are dealing with a very urgent problem for which no effective treatments are known and even small effects could be practically useful.

If inferential statistics are used to detect small effects, further decisions about importance should not exclude the use of the behavior analysis methods. Once inferential statistics are used to decide that an independent variable was probably having some effect on the dependent variable, a prudent researcher would likely consider the magnitudes of effects, their value and costs, and so forth in deciding the practical significance of the effects. Therefore, inferential statistics would not be the sole arbiter of the importance of effects.

Can the use of inferential statistics initially detect small but potentially important effects that are not discoverable with behavior analysis methods? It may be useful to experimentally investigate the relative effectiveness of the two approaches for identifying small effects. Note that arguments favoring inferential statistics for such situations should not assume that the researcher who would examine possible small effects with behavior analysis methods would have no more information than that available in a graphic display of data and would look at the display and render a judgment. Further

note that the complex behavior-analytic methods, sometimes incompletely referred to as visual inspection, often involve the use of a host of aids including drawing straight lines through varying data points, making mathematical calculations to define those straight lines, and examining the distributions of changes in data for individual subjects.

An applied behavior analyst is not likely to rely solely on simple examination of a graphic display if someone's life, health, or welfare depended on a decision about a possible small effect. For example, if we conducted a between-subjects randomized experimental trial and observed the resulting stroke rates for 10,000 people who were given our experimental drug and stroke rates for another 10,000 people not given the drug, we could examine the two resulting distributions of strokes over time. There would likely be great overlap of the two distributions, but we could mathematically characterize the distributions and estimate the likely distribution of effects. If these favored use of the drug, we would proceed to examine all the other considerations that make up the behavior analysis methods. Would we be further interested in whether the effects were statistically significant? We think not. Our best estimate of the distributions of likely future effects is the distribution obtained. Our best estimate of the usefulness of those effects is the result of our use of the behavior analysis methods.

There is a common assumption that inferential statistics provide the advantage of allowing us to estimate or control the likelihood that we will draw incorrect conclusions about effects. Could this be especially useful in the small-effect case? Sedlmeier and Gigerenzer (1989) examined a sample of behavioral research and found that inferential statistics led to incorrect conclusions about effects in fully 60% of the cases (rather than in 5% or 1% of cases, as is commonly assumed, because users of statistics ordinarily op-

erate at the .05 or .01 confidence levels). Abelson (1997) pointed out that Sedlmeier and Gigerenzer's estimate of incorrect conclusions is an aggregated number and that the figure is likely to be higher than that for small-effect cases.

Confidence Intervals or Interval Estimation

Some critics (see, e.g., Cohen, 1990, 1994; Rozeboom, 1960) of the uncritical common uses of inferential statistics in the behavioral sciences have recommended that researchers quit focusing simply on whether an independent variable has a statistically significant effect and instead ask questions about how large that effect might be. Again, suppose we are interested in knowing what effect a training program for mothers had on the number of minutes their children smile each day both before and after the training. Inferential statistics could provide an estimate of an interval or range, which we would expect to include the mean of the increases in minutes of smiling. This would include a probability statement of the likelihood that the mean effect lies within that particular interval.

Even though this approach might appeal in some respects to applied behavior analysts because of our emphasis on estimating magnitudes of effects, we have not seen use of these methods in our literature. Estimating possible magnitudes of effects can be practically important. Perhaps we need to be able to tell mothers how much increase in smiling they could expect as a result of the training program. Our research would not yield a single value for the increase; that value would vary at least over mothers and over days. In other words, instead of a single value for the magnitude of effect, we would obtain a distribution of pre- to posttreatment differences. This distribution would be our best estimate of the effects that resulted from the training program. We could further partition that distribution

to give us a best estimate of the range of effects, the mean effect if we thought that important, and even probability statements (e.g., "This is the smallest effect we have seen, the average increase is 30 minutes per day, and 90% of mothers have obtained at least a 15-minute-per-day increase"). Using inferential statistics to construct confidence intervals seems to add nothing of use to the statements already resulting from our usual methods.

There is one possible use of confidence intervals that we would recommend guarding against. Such methods are usually used as though they allow for the statistical estimation of an interval within which lies the magnitude of the effect of a treatment on behavior. As a sample size increases, inferential statistics would typically yield a progressively narrower interval in which the predicted mean of the increases in smiling would be estimated to lie. There is a tendency in the behavioral sciences to treat such means of effects as the real contribution of the independent variable and the variance around that mean as resulting from chance, all of the unknown and uncontrolled variables that have some effects on the dependent variable thus clouding our view of the true effects of the independent variable. Sidman (1960) warned us against making such assumptions in aggregating data. Our view of behavior is that it is a dynamic interaction with environment that makes the notion of stable, general, single-valued true effects meaningless. Effects of any one independent variable change with a host of other variables. Often many variables simultaneously affect behavior so that we expect differences from subject to subject, observation to observation, and situation to situation. We do not expect to obtain any one true value.

Basic and Conceptually Related Research

This discussion has been aimed primarily towards applied behavioral re-

search, but the same reservations, plus some additional ones, seem to apply to the use of inferential statistics in either basic research or research that is aimed at clarifying conceptualizations. It is unclear to us how simply asking whether an independent variable is likely to have had an effect on a dependent variable (the question most usually addressed by inferential statistics) can be of any importance to either basic or conceptual work. Can it be important to document that another independent variable has some effect, any effect, on a dependent variable? Can any important conceptualization of behavior predict only that a particular independent variable has some effect on a dependent variable? We think not. Inferential statistics speak to questions that seem to be irrelevant to our conceptualizations of the relationships between behavior and the variables with which it interacts.

Perhaps confidence intervals based on inferential statistics could be important to quantitative theories of behavior if the problems of sample size and misleading "true" values could be circumvented. However, such issues do not seem to be foremost in our basic research literature.

Further Considerations

Criticism of the uses of statistics is long-standing. Long before the contemporary use of inferential statistics came into vogue, Bernard (1865/1927) commented on the practice of aggregating data, "even the mathematicians confess that [statistics] can never teach us anything about any particular case. . . . Statistics can therefore bring to birth only conjectural sciences; they can never produce active experimental sciences, i.e., sciences which regulate phenomena according to definite laws" (p. 138). Hogben (1957), a long-term critic of Fisher's (1925) methodology, wrote, "the onus lies on the exponent of statistical theory to furnish irresistible reasons for adopting procedures which have still to prove their worth

against a background of three centuries of progress in scientific discovery accomplished without their aid" (p. 344).

Bakan (1966), a consistent critic of statistical testing of null hypotheses, discussed a number of issues surrounding research and inference in psychology and concluded that "we need to . . . proceed to do investigations and make inferences which bear on them [psychological hypotheses]; instead of, as much of our literature would attest, testing the statistical null hypothesis in any number of contexts in which we have every reason to suppose that it is false in the first place" (p. 436). Cohen (1990), an academic statistician and a major source of the contemporary emphasis on power analysis, observed, "In retrospect, it seems simultaneously quite understandable yet also ridiculous to try to develop theories about human behavior with p [probability] values from Fisherian hypothesis testing and no more than a primitive sense of effect size" (p. 1309).

Paul Meehl (1978), a long-term critic of behavioral science methodology, wrote,

Sir Ronald [Fisher] has befuddled us, and led us down the primrose path. I believe that the almost universal reliance on merely [statistically] refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas [of psychology] is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology. (p. 817)

The complaints against the common uses of statistics are apparently intensifying. As noted above, Sedlmeier and Gigerenzer (1989) have analyzed a sample of research studies and concluded that statistical analyses yielded incorrect decisions about the importance of effects and relationships much more often than the commonly assumed one or five times out of a hundred. Hunter (1997) notes that everyone knows that most null hypotheses are false before the research is even begun. If this is the case, we should assume that most tests of statistical significance will misfire. He calls for a

ban on the use of significance tests for whether independent variables affect dependent variables. Abelson (1997), as also noted above, criticizes as over-generalizations Hunter's and Sedlmeier and Gigerenzer's claims for a 60% error rate in making statistical conclusions, but notes that the error rate is probably far lower than that for research that finds large effects and higher than that for small-effect research. Rosnow and Rosenthal (1988, 1989) have noted the relationship between statistical significance and sample size and have also called for scientists to pay greater attention to effect size. However, they favor the calculation of a product-moment correlation coefficient as an indicant of effect size, an aggregated and indirect estimate. Most of the contemporary critics call for an increasing emphasis on considering magnitudes of effects, usually through interval estimation.

The extreme of the contemporary criticism of the use of inferential statistics has been the actions of the members of a group of psychologists who have asked the American Psychological Association to ban the use of statistical significance tests in research reports published in its journals (see Shrot, 1997).

For a period of time, the editor of the *American Journal of Public Health* banned the use of inferential statistics in testing hypotheses (see DeRouen, 1987; Fleiss, 1986; Lachenbruch et al., 1987; Poole, 1987; Savitz, 1987; Thompson, 1987; Walker, 1986). We do not argue that inferential statistics should be banned from use in published research in applied behavior analysis. Science should be a critical but flexible enterprise. However, given the apparent lack of clear usefulness for inferential statistics, it might be appropriate to require authors to justify their uses of inferential statistics if they wish to publish the results. That requirement might prompt scientists to identify ways in which inferential statistics add value to that resulting from behavior analysis methods in developing, improving, and evaluat-

ing behavioral technologies that are practically important. Similarly, we should remain critical of our use of our more complex data analysis procedures lest they become unimproving and poorly understood rules.

In conclusion, we emphasize that our views of using inferential statistics are somewhat different from those of most of the above-referenced critics. Most of the contemporary complaints against inferential statistics result from the fact that the numerical values of statistics, as shown above, are influenced by sample size. This is the mechanism that yields a high error rate in drawing conclusions and allows the proliferation of misleading conclusions about effects. We have shown that this mechanism necessarily makes the use of inferential statistics a subjective enterprise, contrary to common assumptions. Further, for the kinds of applications for which behavior analysts might be tempted to use inferential statistics, we have argued that those statistics will typically be inferior to the complex methods of behavior analysis and will add nothing to those methods. Finally, the discovery of true effect sizes through interval estimation, the current darling of many of the critics of more traditional statistical significance testing, has been shown to be an artifact of aggregating data that would be better left as individual instances of effects.

REFERENCES

- Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, 8, 12-15.
- Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, 29, 189-194.
- Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis*, 10, 167-172.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Research*, 66, 423-437.
- Barlow, D. H. (Ed.). (1983). *Journal of Applied Behavior Analysis*, 16(2, 3).
- Bernard, C. (1927). *An introduction to the study of experimental medicine*. New York: Macmillan. (Original work published 1865)
- Branch, M. N. (Ed.). (1993). *Journal of the Experimental Analysis of Behavior*, 60.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304-1312.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- Cooper, J. O., Heron, T. E., & Heward, W. L. (1987). *Applied behavior analysis*. Columbus, OH: Merrill.
- Costa, L., & Rourke, B. P. (Eds.). (1994). *Journal of Clinical and Experimental Neuropsychology*, 16(1).
- DeRouen, T. A. (1987). Comment on statistical testing and confidence intervals. *American Journal of Public Health*, 77, 237.
- Doob, L. W. (Ed.). (1994). *Journal of Social Psychology*, 134(4).
- Dugbartey, A. T. (1994). The factor structure of traditional beliefs among Ghanaian university students. *Journal of Social Psychology*, 134, 549-550.
- Edgington, E. S. (1980). Random assignment and statistical tests for one-subject experiments. *Behavioral Assessment*, 2, 19-28.
- Edgington, E. S. (1982). Non-parametric tests for single-subject multiple schedule experiments. *Behavioral Assessment*, 4, 83-91.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Boyd.
- Fleiss, J. L. (1986). Significance tests have a role in epidemiologic research: Reactions to A. M. Walker. *American Journal of Public Health*, 76, 559-560.
- Gentile, J. R., Roden, A. H., & Klein, R. D. (1972). An analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, 5, 193-198.
- Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.
- Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge, England: Cambridge University Press.
- Gottman, J. M., & Glass, G. V. (1978). Analysis of interrupted time-series experiments. In T. R. Kratochwill (Ed.), *Single-subject research: Strategies for evaluating change* (pp. 197-235). New York: Academic Press.
- Hartmann, D. P. (1974). Forcing square pegs into round holes: Some comments on "An analysis of variance model for the intrasubject replication design." *Journal of Applied Behavior Analysis*, 7, 635-638.
- Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis*, 13, 543-559.
- Hays, W. L. (1963). *Statistics for psychologists*. New York: Holt, Rinehart, and Winston.
- Hogben, L. (1957). *Statistical theory: The re-*

- relationship of probability, credibility, and error: An examination of the contemporary crisis in statistical theory from a behaviorist viewpoint. London: Allen and Unwin.
- Horne, G. P., Yang, M. C. K., & Ware, W. B. (1982). Time-series analysis for single-subject designs. *Psychological Bulletin*, 91, 178–189.
- Howell, D. C. (1992). *Statistical methods for psychology* (3rd ed.). Boston: PWS-Kent.
- Huitema, B. E. (1986). Statistical analysis and single-subject designs. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis* (pp. 209–232). New York: Plenum.
- Hulse, S. H. (Ed.). (1994). *Journal of Experimental Psychology: Animal Behavior Processes*, 20(3).
- Hunt, E. (Ed.). (1994). *Journal of Experimental Psychology: General*, 123(2).
- Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8, 3–7.
- Hurlburt, R. T. (1994). *Comprehending behavioral statistics*. Pacific Grove, CA: Brooks/Cole.
- Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Jones, R. R., Vaught, R. S., & Weinrott, M. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis*, 10, 151–166.
- Kazdin, A. E. (1976). Statistical analyses for single-case experimental designs. In M. Hersen & D. H. Barlow (Eds.), *Single-case experimental designs: Strategies for studying behavior change* (pp. 265–316). Oxford, England: Pergamon Press.
- Keselman, H. J., & Leventhal, L. (1974). Concerning the statistical procedures enumerated by Gentile et al.: Another perspective. *Journal of Applied Behavior Analysis*, 7, 643–645.
- Kratochwill, T. (Ed.). (1974). *Single-subject research: Strategies for evaluating change*. New York: Academic Press.
- Kratochwill, T., Alden, K., Demuth, D., Dawson, D., Panicucci, C., Arnston, P., McMurray, N., Hemstead, J., & Levin, J. (1974). A further consideration in the application of an analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, 7, 629–633.
- Kratochwill, T., & Levin, J. R. (1980). On the applicability of various data analysis procedures to the simultaneous and alternating treatment designs in behavior therapy research. *Behavioral Assessment*, 2, 353–360.
- Lachenbruch, P. A., Clark, V. A., Cumberland, W. G., Chang, P. C., Afifi, A. A., Flack, V. F., & Elashoff, R. M. (1987). Comment on statistical testing and confidence intervals. *American Journal of Public Health*, 77, 237.
- Levin, J. R. (Ed.). (1994). *Journal of Educational Psychology*, 86(3).
- Mainstone, L. E., & Levi, A. S. (1987). Fundamentals of statistical process control. *Journal of Organizational Behavior Management*, 9, 5–21.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–834.
- Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis*, 7, 647–653.
- Neef, N. A. (Ed.). (1993). *Journal of Applied Behavior Analysis*, 26(2, 3).
- Nevin, J. A. (Ed.). (1983). *Journal of the Experimental Analysis of Behavior*, 40.
- Notz, W. W., Boschman, I., & Tax, S. T. (1987). Reinforcing punishment and extinguishing reward: On the folly of OBM without SPC. *Journal of Organizational Behavior Management*, 9, 33–46.
- Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis* (pp. 157–186). New York: Plenum.
- Pfadt, A., Cohen, I. L., Sudhalter, V., Romanczyk, R. G., & Wheeler, D. J. (1992). Applying statistical process control to clinical data: An illustration. *Journal of Applied Behavior Analysis*, 25, 551–560.
- Pfadt, A., & Wheeler, D. J. (1995). Using statistical process control to make data-based clinical decisions. *Journal of Applied Behavior Analysis*, 28, 349–370.
- Poole, C. (1987). Beyond the confidence interval. *American Journal of Public Health*, 77, 195–199.
- Rayner, K. (Ed.). (1994). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(5).
- Rosnow, R. L., & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. *Journal of Counseling Psychology*, 35, 203–208.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276–1284.
- Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57, 416–428.
- Savitz, D. (1987). Comment on statistical testing and confidence intervals. *American Journal of Public Health*, 77, 237–238.
- Schmitt, N. (Ed.). (1994). *Journal of Applied Psychology*, 79(3).
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316.
- Shrot, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, 8, 1–2.
- Sidman, M. (1960). *Tactics of scientific research*. New York: Basic Books.

- Somerville, S. C. (Ed.). (1994). *Child Development* 65(3).
- Spence, J. T., Cotton, J. W., Underwood, B. J., & Duncan, C. P. (1990). *Elementary statistics* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Thompson, W. D. (1987). Statistical criteria in the interpretation of epidemiologic data. *American Journal of Public Health*, 77, 191–194.
- Thoresen, C. E., & Elashoff, J. D. (1974). "An analysis-of-variance model for intrasubject replication design": Some additional comments. *Journal of Applied Behavior Analysis*, 7, 639–641.
- Van Meter, P., Yokoi, L., & Pressley, M. (1994). College students' theory of note taking derived from their perceptions of note-taking. *Journal of Educational Psychology*, 86, 323–338.
- Walker, A. M. (1986). Reporting the results of epidemiologic studies. *American Journal of Public Health*, 76, 556–558.
- Williams, D. G. (1994). Population density and mental illness. *Journal of Social Psychology*, 134, 545–546.
- Wolff, A. B., Sass, K. J., & Keidan, J. (1994). Case report of an intracarotid amobarbital procedure performed for a deaf patient. *Journal of Clinical and Experimental Neuropsychology*, 16, 15–20.