The Behavior Analyst 1999, 22, 83–85 No. 2 (Fall)

Statistical Inference in Behavior Analysis: Friend or Foe?

Alan Baron
University of Wisconsin–Milwaukee

Behavior analysts are undecided about the proper role to be played by inferential statistics in behavioral research. The traditional view, as expressed in Sidman's *Tactics of Scientific Research* (1960), was that inferential statistics has no place within a science that focuses on the steady-state behavior of individual organisms. Despite this admonition, there have been steady inroads of statistical techniques into behavior analysis since then, as evidenced by publications in the *Journal of the Experimental Analysis of Behavior*. The issues raised by these developments were considered at a panel held at the 24th annual convention of the Association for Behavior Analysis, Orlando, Florida (May, 1998). The proceedings are reported in this and the following articles.

Key words: research design, statistical inference, experimental control, single-subject design

This and the following five articles report the proceedings of a panel held at the 24th annual convention of the Association for Behavior Analysis in Orlando, Florida. The panel members, besides myself, were Nancy Ator, Marc Branch, John Crosbie, Michael Davison, and Michael Perone. Subsequent to the meeting, Richard Shull prepared a commentary based on written versions of the panelists' presentations. His comments are included as a sixth article. As chair of the panel, I introduced the topic:

Behavior analysis has had a stormy relationship with statistical methods. Note that our concern is with those statistical procedures referred to as "inferential," in other words, procedures whose goal is to infer characteristics of a hypothetical population from a limited set of sample observations. There is no quarrel about the value of descriptive statistics. Experimental data must be organized and summarized, and the methods of behavioral analysts are similar to those of workers in other research traditions. Thus, we too use the mean as a measure of central tendency, the standard deviation as a measure of variation, and regression analysis as a measure of association.

The issues first appear to have been broached in *The Behavior of Organisms,* published in 1938, where B. F. Skinner laid out the principles with which we are familiar. His research had a number of unique features, not the least of which was the unconventional way that he designed his experiments and treated the results.

Unlike much of the research of that time, Skinner's experiments focused on the behavior of individual subjects. He studied a small number of rats under a series of conditions, and within each condition observations were prolonged until the animal's performance was stable. This was quite different from the usual procedure in which groups of subjects were observed briefly, one experimental condition per group, and the results were reported as the average performance for the group. Also different was the way in which Skinner analyzed the results. Each subject's performance was reported and conclusions were reached from data summarized in graphs, usually cumulative records. He saw no need to verify his re-

This and the following papers were presented at the 24th annual convention of the Association for Behavior Analysis, Orlando, Florida, May, 1998. I thank the panel members, Nancy Ator, Marc Branch, John Crosbie, Michael Davison, and Michael Perone, and the commentator, Richard Shull, for helping me to understand better the place of inferential statistics in behavior analysis.

Please address correspondence about this article to Alan Baron, Department of Psychology, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin 53201 (E-mail: ab@uwm.edu).

TABLE 1

Questions to address

Concerning your own research:
Do you use inferential statistical methods to analyze the results?
If so, how do you decide when statistical verification of findings is needed?
If not, how do you establish the reliability of your findings?
Concerning research done by others:
Do you regard findings from steady-state and group-statistical methods as interchangeable?
If not, how do you reconcile experiments that use different methods?
Concerning graduate education and publication policies:
Should training in statistics by regarded as essential preparation for a career in the experimental analysis of behavior?
What should the stance of the *Journal of the Experimental Analysis of Behavior* be about using statistics to clarify experimental results?

sults through the statistical tests that were coming into vogue at that time.

In *The Behavior of Organisms,* Skinner did not explain or defend his approach until the very last few pages when almost as an afterthought he briefly gave his reasons for deviating from more conventional methods. Later on, Murray Sidman discussed these reasons in considerably more detail in what has become the definitive methodological treatise for behavior analysis, *Tactics of Scientific Research* (1960).

According to Skinner and Sidman, group experiments and the statistical methods used to analyze the results are seriously flawed. The limitations are familiar to students of behavior analysis. First, although group differences may be statistically significant, the group averages conceal exceptions at the individual level. Laws that apply only to groups do not allow prediction of what the single organism will do. Second, when functional relationships are derived from groups exposed to different treatments, the functions do not originate in the behavior of any particular organism. In other words, the average performances from different subjects do not have a clear counterpart in nature. Finally, and most troubling, is that the group-statistical method diverts the researcher from a full experimental analysis. The hallmark of the scientific experiment is the manipulation and control of variables. But the group-statistical researcher must be prepared to tolerate uncontrolled differences among subjects as an inherent feature of the research. This concession appears to undermine the reason for conducting experiments in the first place.

By comparison, Skinner's single-subject designs do not have these problems. The inquiry focuses on the ultimate concern: the behavior of the individual organism. Extraneous variables are controlled within the experiment rather than averaged out statistically. Effects of variables—functional relationships—are examined as they naturally occur, within the same organism rather than as a construction from the performance of a group. And there is no need for inferential statistics because behavior is observed as a steady state.

To judge from the research literature, the methods of Skinner and Sidman were taken quite seriously by the behavior-analytic researchers of that time. Several years ago, I surveyed the designs of experiments published in issues of the *Journal of the Experimental Analysis of Behavior* during the late 1950s (Baron, 1990). Not surprisingly, inferential statistics were almost never used. They were employed in fewer than 5% of the reports, and it was the rare experiment that followed a between-group design (the type of design most closely linked to inferential tests). But over the years, something happened. By 1990, close to one third of the experiments employed inferential statistics, and between-group designs also had increased in frequency. Although I do not have any new data to present today, my impression is that this trend is continuing.

So it appears that a significant number of us are violating the precepts of

our founding fathers. The following six articles address this puzzling discontinuity between contemporary behavior-analytic research practices and traditional views of the way that research should be conducted. The individuals I enlisted in this effort all have distinguished research records, as well as considerable experience as editors and reviewers of behavior-analytic research. To organize our discussion, I provided them with a series of questions (see Table 1). However, knowing that they all are free spirits, I will not be offended if they take the discussion in other directions; our goal is to explore the issues as broadly as we can.

REFERENCES

Baron, A. (1990). Experimental designs. *The Behavior Analyst, 13,* 167–171.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* New York: Basic Books.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.* New York: Appleton-Century-Crofts.

The Behavior Analyst 1999, 22, 87–92 No. 2 (Fall)

Statistical Inference in Behavior Analysis: Some Things Significance Testing Does and Does Not Do

Marc N. Branch
University of Florida

Significance testing plays a prominent role in behavioral science, but its value is frequently overestimated. It does not estimate the reliability of a finding, it does not yield a probability that results are due to chance, nor does it usually answer an important question. In behavioral science it can limit the reasons for doing experiments, reduce scientific responsibility, and emphasize population parameters at the expense of behavior. It can, and usually does, lead to a poor approach to theory testing, and it can also, in behavior-analytic experiments, discount reliability of data. At best, statistical significance is an ancillary aspect of a set of data, and therefore should play a relatively minor role in advancing a science of behavior.

One need only look at a few scientific journals in the domain of behavioral science to note the ubiquity of statistical significance testing (cf. Hubbard, Parsa, & Luthy, 1997; Sterling, 1959). Such information can be found even in journals that emphasize replication and individual-subject data (Hopkins, Cole, & Mason, 1998). It is interesting that behavioral science has aligned itself so strongly with significance testing, whereas other, more successful sciences have not. Significance testing has been ubiquitous in psychology since the early 1950s, yet it is difficult to discern how its use has improved the field. Many important behavioral processes were discovered and analyzed prior to the development and implementation of tests of statistical significance, and virtually all of modern chemistry and physics was developed without their assistance. It is clear, then, that science can proceed without significance testing, and it is not at all clear that significance testing has helped to advance behavioral science. The mystery is why so many behavioral scientists continue to place such high value on the methods of significance testing.

Many journals require tests of statistical significance as part of the data analyses. Given that it is possible to conduct good research without resorting to this method, it seems unwise for journals to have editorial policies that make such tests necessary for publication. Nevertheless, such policies exist, and that presents a problem for the researcher who prefers not to employ the tests. In the comments that follow, I hope to help researchers who find themselves at the mercy of journal editors or grant reviewers who clamor for significance tests use the old maxim, "Sometimes the best defense is a good offense." Specifically, I'll try to point to both inherent weaknesses in the logic of significance testing and to potential consequences of unreasonable allegiance to them. These points may be used in interchanges with editors and reviewers to illustrate that preference for not using significance testing is fully defensible. Virtually none of the weaknesses and negative consequences of significance testing that I shall present are new or recently discovered. All have been known and discussed many times over the years. My comments, then, should stand only as reminders.

Preparation of the paper was assisted by USPHS Grant DA-04074 from the National Institute on Drug Abuse.

Reprints can be obtained from the author at the Psychology Department, University of Florida, Gainesville, Florida 32611 (E-mail: branch@psych.ufl.edu).

THINGS THAT SIGNIFICANCE TESTING DOES NOT DO

Tests of Statistical Significance Do Not Provide a Quantitative Estimate of the Reliability of a Result

This fact means that the expression "statistically reliable" is a non sequitur. There is no dispute about this issue. There is no known relationship between level of significance and replicability (see Carver, 1978). A p value, or its inverse, therefore is not an estimate of how likely the results obtained are to be replicated. That is, to say that $p < .001$ does not imply that there is only one chance in a thousand that a replication would fail, nor does it mean if you conducted the experiment 1,000 times, you would most likely find only one discrepant result. The only way currently known to determine if a finding is reliable is to replicate the result. How many replications must be performed before we agree that a result is reliable? The answer is, "It depends." And it depends on many things. The number of replications needed to be convincing depends, for example, on what is already known. As an illustration, if I throw a brick at a pane of glass and it breaks, most people would not ask that I replicate the effect to be sure to conclude that the brick hitting the glass was the cause of the glass shattering. (Nor would most people ask for a "control" pane of glass, but that is another, although related, issue.) Why not? Because of what we already know about glass, bricks, and thrown objects. Similarly, in science, some things, because we know relatively little, require more extensive replication, whereas other kinds of results do not require as much (these are more likely in physics, however, than in behavioral science).

A p value has quantitative meaning only if the null hypothesis is true. Of course, we do not know if the null hypothesis is true. That is why we're doing the test in the first place—to get information to help us decide if it is true. If we knew that the null hypothesis were true, then a p value of .001 would indicate that if we perfectly replicated the study 1,000 times, we should expect only one case to come out differently. Of course, if we *knew* the null hypothesis were true, we would have no reason to do the test.

Tests of Statistical Significance Do Not Estimate the Probability That the Results Were Due to Chance

This fact can be illustrated easily by remembering that a p value is a conditional probability. Specifically, a p value represents the probability of a certain kind of data given that the null hypothesis is true: $p = p(\text{data}|H_0)$. To state that something is due to chance is to reverse the conditionality. That is, saying that some result is due to chance is essentially stating that a p value is an estimate that the null hypothesis was operating, and that simply is not the case because that would imply that $p = p(H_0|\text{data})$. Elementary probability theory tells us that $p(A|B)$ is not equal to $p(B|A)$, except in the rare case that the probabilities are independent of one another and are also equal (Parzen, 1960). It is easy to illustrate to oneself that they are not equal by considering everyday examples like $p(\text{raining}|\text{cloudy})$ versus $p(\text{cloudy}|\text{raining})$ or $p(\text{manuscript rejected}|\text{submitted to } \textit{The Behavior Analyst})$ versus $p(\text{submitted to the } \textit{The Behavior Analyst}|\text{manuscript rejected})$. All this is to say that p is not an estimate of the probability of the truth of the null hypothesis. That is one of the important reasons why it is not an estimate of the reliability of results.

Tests of Statistical Significance Usually Do Not Answer a Question to Which the Answer Is Unknown

Significance tests provide researchers with evidence on which to decide if the null hypothesis is true. That is not a very important question to answer, because in the vast majority of cases the null hypothesis of no differ-

ence is not true. As Meehl (1967) notes, he and Lykken have shown that to be the case with data. When one considers what the null hypothesis usually entails (i.e., that the only thing operating in the experiment is randomness), it is pretty obvious that there are very few situations in which that could be true. As noted by Kraemer (1998), "something nonrandom is almost always going on, and it seems a trivial exercise to redemonstrate that fact" (p. 206). Virtually all behavioral scientists learn that by increasing *N* one increases the likelihood of finding statistical significance, but somehow the implications of that fact get lost. If it is so important, how come it is so easy to influence? One can relate this point to the previous two. Perhaps it is not so bad that significance tests do not estimate the truth of the null hypothesis, because we already know that it is false. Also, a procedure designed to help us decide about something we already know could hardly be one that would provide quantitative estimation of reliability.

It seems to me that the three arguments just presented would be sufficient to convince a reasonable person that eschewing significance testing for other approaches to establishing reliability is a perfectly sane and defensible position. Perhaps, however, one might be faced with someone who is less than reasonable and therefore might want to use the following points about unfortunate consequences of null-hypothesis significance testing.

THINGS THAT SIGNIFICANCE TESTING DOES DO

Tests of Statistical Significance Reduce Scientific Responsibility

This point is made eloquently by Carver (1978), who points out that slavish adherence to significance testing is a view in which a scientist is given full responsibility for the origin, design, and conduct of an experiment, but is given no responsibility for deciding whether the results are useful or meaningful. The thing that sets science apart as a social enterprise is that it is self-corrective. The mechanism of correction is replication. It is through replication that confidence in a finding is established, and it is through failures of replication that mistakes are corrected. Social contingencies are important in science, and significance testing blunts their effectiveness. Before tests of significance were invented, a scientist's reputation depended on the reliability of his or her descriptions of results and conclusions drawn from them. If a scientist claimed to observe some result, and subsequently it was shown that the result was not reliable, the scientist's reputation suffered. With significance testing, there is an out. For example, suppose Scientist A performs an experiment and gets a statistically significant result and makes claims on that basis. Other scientists perform replications but do not get the same result. Does Scientist A's reputation suffer? No, because he or she can claim, "It's not my fault. We expect some proportion of errors when using tests of statistical significance, and this was one of them. I played by the rules and am therefore blameless." If one's reputation rides on the reliability of findings, you can bet that scientists would be more careful about what they publish.

The fact that tests of statistical significance protect a scientist to some degree may be one of the factors that determine their popularity. Use of significance testing might be thought of as a form of avoidance, avoidance of social censure.

Requiring statistical significance as a prerequisite for publication also serves to blunt science's most precious resource for ferreting out mistakes (or even fraud). As noted above, the self-corrective nature of science is based on replication. Failures to replicate are very, very important in informing us that we don't fully understand what is going on (Sidman, 1960). If statistical significance is requisite for publication, then one will have a difficult time pub-

lishing results of experiments in which replication of a statistically significant effect fails.

Tests of Statistical Significance Are Frequently Employed in a Poor Manner to Test Theory

I like to call this the "dumb-null-hypothesis problem," and it is a point well made by Meehl (1967, 1978). Consider the usual arrangement for using statistical significance testing to put a theory to test. The null hypothesis is set at "no effect." The alternative hypothesis, the one that the theory predicts, is set at "some effect." If a statistically significant effect is obtained, the null hypothesis is rejected, and the alternative hypothesis, and therefore the theory, gain support. Consider now how statistical significance is determined. A statistic that is a ratio of "effect variance" over "error variance" is computed. If that ratio is large enough, statistical significance is achieved. Next, consider the effects of improvement in experimental technique. Better control of extraneous variables should decrease error variance, and therefore make it easier for the ratio to reach the critical value. Thus, as methods are refined and better experiments are conducted, it becomes easier to demonstrate statistical significance. That means that better methods make it easier to reject the null hypothesis and therefore support the theory, no matter what the theory is. Obviously, this is not a very good outcome.

There is a way to circumvent this issue and make use of significance testing in a more rational fashion. Statistics courses usually inform us that it is not necessary that the null hypothesis be "no effect." Instead it can be set at some particular effect; that is, it can be set at what the theory predicts. Then, as experimental techniques are improved it still becomes easier to reject the null hypothesis, but in this case the null hypothesis is what the theory predicts. Therefore, as experimental techniques improve, the theory is put to a more severe test, exactly the kind of result for which one would hope. This latter approach is exactly what is done in curve fitting (Lewis, 1966), a strategy favored by the more advanced sciences.

Tests of Statistical Significance Emphasize Population Parameters Over Behavior

As Danziger (1987, 1990) has noted, a remarkable development in behavioral science in the last half of this century has been the emergence of the aggregate as the unit of analysis. This is an odd development in a field presumably dedicated to understanding behavior or "the mind." Behavior is something an individual does, not what a group average does. (It is especially difficult to think of "group mind.") The direction of inference from a group average is to the population, not to the individual, so when the unit of analysis becomes the aggregate we develop a science not of behavior but of population parameters. Perhaps some take comfort from the view that something that provides information about the population is inherently more general than something that applies to some individuals, but that comfort ought to be tempered by the realization that the generality is not about behavior. A 2% rate of pregnancy in a population may have important meaning for policy makers (insurance and public), but it has no meaning for an individual female, who is never 2% pregnant.

Tests of Statistical Significance Can Limit the Reasons for Doing Experiments

Tests of statistical significance generally require that the experimental question be a test of a hypothesis. Testing hypotheses, of course, is an honored tradition in science and certainly a worthy enterprise. As ably noted by Sidman (1960), however, there are many other very good reasons for doing experiments. Isaac Newton, a sci-

entist of some note and success, suggested that one should never have a hypothesis, but rather should simply ask questions about Nature. His successes make clear that hypothesis testing is not the only route to achievement in science. Having to shoehorn one's experiments into the logic of hypothesis testing frequently leads to absurdities like developing one's "hypotheses" after the data are collected.

Tests of Statistical Significance Often Discount Reliability in Effects in the Case of Behavior-Analytic Experiments

This happens because the tests typically ignore the replications inherent in behavior-analytic research designs. Consider, for example, cases in which for each subject a stable baseline is established in each condition of the experiment (a very common occurrence in such research). Statistical analyses will then proceed using, for example, averages from the last five sessions of observation in each condition. This mean is treated as if it were a single score from a single observation, but it clearly is not. Each of the last five sessions of each condition constitutes a replication, so in reality, at a minimum, five replications of the value are ignored. Given other evidence of good experimental control, five replications of a value provide direct information about the reliability of the value, and that information is lost in the statistical test. The expression, "at a minimum," was used above because in many cases the session average itself underestimates the reliability of the measure. Suppose that in each session a variable-interval schedule was in effect, and cumulative records reveal that rate of behavior was constant throughout the session. If the rate of behavior is reported as the session average, this single number does not tell us as much about reliability of the effect. A cumulative record, however, reveals that from minute to minute the effects were reliable. I know of no statistical test that can deal with that kind of reliability.

Given that tests of statistical significance, despite any evidence that they have assisted the development of behavioral science, have become such an integral feature of research in behavioral science, it seems highly unlikely that we shall at any time soon see a broad deemphasis of their use. There are rumblings, however, in the psychological sciences that may indicate that slavish attachment to significance testing may eventually fade away (e.g., Cohen, 1994; Hunter, 1997; Loftus, 1996). If that is the case, research conducted in the tradition of behavior analysis, research that is directed at individual behaving subjects and that employs methods that directly illustrate reliability, can serve as a model for other researchers. Now could well be a very opportune time for behavior analysis, a time in which behavior analysis illuminates the way for other researchers in psychology. Behavior analysts should not "hitch themselves to a wagon from which people are leaping," but instead should look upon the next decade as one in which behavior-analytic methods gain even wider popularity.

REFERENCES

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Danziger, K. (1987). Statistical method and the historical development of research practice in American psychology. In L. Kruger, G. Gigerenzer, & M. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 35–47). Boston: MIT Press.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research.* New York: Cambridge University Press.

Hopkins, B. L., Cole, B. L., & Mason, T. L. (1998). A critique of the usefulness of inferential statistics in applied behavior analysis. *The Behavior Analyst, 21,* 125–137.

Hubbard, R., Parsa, R. A., & Luthy, M. R. (1997). The spread of statistical significance testing in psychology: The case of the *Journal of Applied Psychology,* 1917–1994. *Theory and Psychology, 7,* 545–554.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 307.

Kraemer, H. C. (1998). Statistical significance: A statistician's view. *Behavioral and Brain Sciences, 21,* 206–207.

Lewis, D. (1966). *Quantitative methods in psychology.* Iowa City: University of Iowa Press.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

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.

Parzen, E. (1960). *Modern probability theory and its applications.* New York: Wiley.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

The Behavior Analyst 1999, 22, 93–97 No. 2 (Fall)

Statistical Inference in Behavior Analysis: Environmental Determinants?

Nancy A. Ator
Johns Hopkins School of Medicine

Use of inferential statistics should be based on the experimental question, the nature of the design, and the nature of the data. A hallmark of single-subject designs is that such statistics should not be required to determine whether the data answer the experimental question. Yet inferential statistics are being included more often in papers that purport to present data relevant to the behavior of individual organisms. The reasons for this too often seem to be extrinsic to the experimental analysis of behavior. They include lapses in experimental design and social pressure from colleagues who are unfamiliar with single-subject research. Regardless of whether inferential statistics are used, behavior analysts need to be sophisticated about experimental design and inferential statistics. Such sophistication not only will enhance design and analysis of behavioral experiments, but also will make behavior analysts more persuasive in presenting rationales for the use or nonuse of inferential statistics to the larger scientific community.

I am by no means a sophisticated statistician and never liked math. In fact, it was a source of great relief as a graduate student to realize that my chosen enclave of research in psychology not only did not use inferential statistics but had well-founded and eloquently stated conceptual reasons for eschewing both them and the realm of "hypothesis testing" itself (Sidman, 1960; Skinner, 1950). During my tenure as an associate editor of the *Journal of the Experimental Analysis of Behavior* (*JEAB*), I found, with some dismay, that people were submitting papers that included statistical analyses to *JEAB*, the bastion of single-subject design in basic research. I had long since faced the fact that I did need to learn something about statistics and had even included an analysis of variance (ANOVA) or two in my own manuscripts; but the editorial responsibilities provided the impetus to really think about the issues of *why?*, *when?*, and *which?* I began looking at statistics in all manuscripts (not just *JEAB* manuscripts) in a different light and listening to colleagues talk about their approaches to data analysis from a new perspective.

As a light-hearted summary of what I have seen and heard, I present below the top 10 reasons given for using inferential statistics:

10. "My experiment used a truly randomized design."

9. "I couldn't use a truly randomized design for practical reasons, but I planned subject assignment in advance to compensate for nonequivalence of groups."

8. "I'm doing research in a clinical setting. I plan to do a time-series analysis with my single-case data, because I have limited flexibility in conducting reversals and manipulating parameters of the treatment."

7. "I'm doing single-subject research with college students because I won't need as many as a group design; but there's no way I can run all the conditions to stability or do replications and be finished by the end of the semester. So, I figure I can do an ANOVA on the group data as a back-up to the individual graphs."

6. "I thought I was doing a single-subject design with my rat study, but I wasn't able to keep up with the data

Preparation of this manuscript was supported by Grant RO1-DA04133 from the National Institute on Drug Abuse.

Reprints can be obtained from the author at the Behavioral Biology Research Center, Johns Hopkins Bayview Campus, 5510 Nathan Shock Drive, Suite 3000, Baltimore, Maryland 21224-6823 (E-mail: ator@mail.jhmi.edu).

before running tests, and it turns out the baselines were really different across rats, and I can't tell if there's anything there, and I need to get a publication out of this!"

5. "I'm an assistant professor working towards tenure. Even though I'm committed to an *experimental* analysis of behavior, I need to throw in some statistics because a senior faculty member gave me a really hard time at a departmental seminar about how I could make anything out of so few subjects."

4. "Look, my last grant application got shot down in study section because I didn't include any plan for inferential statistics. I can't afford to let that happen again."

3. "My last manuscript got shot down by a reviewer who wasn't convinced that I had a reliable effect and wanted to see some statistics."

2. "I really think these single-subject data are best suited for Journal XYZ, but I hear the new editor is biased against papers that don't include statistics."

1. "The guy down the hall got this great software package that gave me a $p < .001$, so I think I'll include it!"

These reasons can be separated easily into those that are legitimate and those that are less so. They can be categorized as ones for which the rationale for using inferential statistics is intrinsic to the nature and design of the research and those for which the rationale is extrinsic to the experimental question. In the remainder of the paper, I will discuss the issues raised in the list above. I will conclude with what seems to me to be the antidote to extrinsic determinants of the use of inferential statistics in the experimental analysis of behavior. (I must admit that some of these extrinsic reasons have affected even my own behavior over the years.)

INTRINSIC REASONS FOR USING STATISTICS

Inferential statistics are, of course, appropriate for true group designs, that is, for experiments that use random assignment of subjects to conditions or conditions to subjects, and plan to control for variability via statistical methods. Too, there are procedures for handling research situations in which subjects cannot be assigned randomly so that inferential statistics still are appropriate. Reasons 10 and 9, which describe classical group designs, are, of course, recognized as necessary conditions for use of inferential statistics (Bordens & Abbott, 1996; Rosnow & Rosenthal, 1996).

Reason 8, which refers to behavioral treatment research, also is a legitimate rationale for inferential statistics. Within the world of single-subject or single-case designs, there are appropriate statistical methods to aid evaluation of treatments in which conditions, usually in clinical settings, are not optimal for experimental control. Texts on statistics for single-case designs discuss the problems and pitfalls of such research (e.g., limitations on reversals or the ability to manipulate parameters). They set forth ways in which statistics can help researchers draw appropriate conclusions from data collected in settings in which true *experimental* analysis is not possible (Bordens & Abbott, 1996; Kazdin, 1984; Krishef, 1991).

EXTRINSIC REASONS FOR USING STATISTICS

Reasons 7 through 1 are problematic. Whether you agree with behavior-analytic colleagues who see a useful role for inferential statistics within the experimental analysis of behavior or not, these seven reasons are extrinsic to sound science.

Faulty Planning, Real Life, and the Desire to Publish

The art of experimental design, whether group or single-subject, sometimes seems to be falling by the wayside. The publication manual of the American Psychological Association used to have "Design" as the first section of "Methods"; now the term is not

even in the index. Reasons 7 (research using college students) and 6 (research using rats) exemplify those situations in which experiments are not planned with an eye to the strongest possible design for the experimental question.

The single-subject design has a practical appeal over group designs because it requires fewer subjects, does not require randomization, and discourages the practice of gathering pilot data (Sidman, 1960). So, it is sometimes too easy to get started on an experiment—perhaps even in the spirit of Skinner's (1956) famous first unformalized principle of scientific practice, "When you run onto something interesting, drop everything else and study it."

The rub comes when one is faced with the labor and time commitment involved in *experimental* analysis of behavior: stability criteria, the number of reversals needed to manipulate an independent variable, equating baselines across subjects, parametric manipulations, and close monitoring for long periods of time. Then, complications occur: aging rats, equipment that malfunctions, baselines that drift, unexpected variability across subjects, and unexpected order effects. All this labor and these complications occur in the context of academic deadlines, funding deadlines, promotion and tenure reviews, and other realities of life.

It is no wonder that there is great appeal in taking what data have been collected, putting them through a few statistical manipulations, and seeing whether there is "anything there." Some of these manipulations *can* turn out pretty well, and, regardless of the design, well-collected, interesting, and clear data should be published. Sometimes though, the result is neither fish nor fowl—a hybrid single-subject/group design with the best of neither: few subjects, great variability, little replication, few parametrics, mean data that are "significant" but represent few if any of the subjects, analyzed with statistical procedures that are arguably inappropriate for the data. Sometimes, this approach has been encouraged another class of environmental det minants: the academic social envirc ment.

The Devil Made Me Do It

Reasons 5 through 2 describe s cially mediated contingencies that su port inclusion of statistical analyse As behavior analysts make their w in the academic world, it is a fact life that at one time or another, our r search presentations are questioned f their lack of p values. Adding statisti then becomes an avoidance respons that heads off criticism from co leagues who refuse to take seriousl any result not accompanied by "$p <$ .05."

When behavior analysts do not un derstand design and statistical method well enough, we become unduly sub ject to the preconceived notions of re viewers, editors, colleagues, and de partment chairs, who are not trained t appreciate steady-state research. B unduly subject, I mean that we canno stand up for single-subject designs in a credible way. Many of the very people (reviewers and editors) who call for more statistics do not understand them either, and the contingency seems to be placed on having a p value. To the extent this is true, many of the statistics reported in psychological journals seem to represent rule-governed behavior run amok! To be able to argue effectively against such rule-governed behavior, behavior analysts who believe inferential statistics to be inappropriate for their data must gain a thorough understanding of how such statistics should be used and what they can and cannot do (Branch, 1999; Perone, 1999).

Beware the Gold Star

Reason 1, the great p value provided by statistical analysis software, while the most facetious, may be the most insidious. Because good steady-state research with lots of control of environmental variables (not to mention the

"salutary" influence of autocorrelation) produces results with clear differences in effects, it turns out to be remarkably easy to find significant p values even with few subjects. With easy-to-use statistical software, who can fail to enjoy the immediate reinforcement of plugging data into a spreadsheet and, in the wink of an eye and the click of a mouse, seeing "$p = .0001$." Like pasting a gold star on your data! The emergence of symposia, journal articles, and a task force questioning the unwarranted "significance" of p values, however, should suggest caution (cf. Harris, 1997; Hopkins, Cole, & Mason, 1998; Loftus, 1996). The wind may be changing.

CONCLUSION

Rather than make an appeal for or against a role for statistics in behavior analysis, I want to make an appeal for thoughtfulness in experimental design and for being more sophisticated in our knowledge of statistics. Use of inferential statistics should be based on the experimental question, the nature of the design, and the nature of the data.

A hallmark of single-subject designs in the experimental analysis of behavior is that such statistics should not be required to determine the reality of the effect of an independent variable. Although there are situations in which inferential statistics can be useful adjuncts to visual inspection of the data in single-subject designs (Fisch, 1998; Krishef, 1991), the use of statistics for reasons that are extrinsic to the experimental design should be minimized. Sometimes use of statistics is the result of poor planning or execution of an experiment: a quasi-single-subject design. Although efforts to salvage data that have been carefully collected are defensible, the statistics as used often are not.

Behavioral science can only be strengthened by a decline in the kind of social variables suggested in Reasons 7 through 1 as primary determinants of using a p value. To counteract these social influences, students of the experimental analysis of behavior should be taught statistics as thoroughly as possible, and the rest of us should brush up. Behavior analysts need to be sufficiently sophisticated about the experimental designs they use to be able to argue persuasively for the most appropriate analysis of the data, given the design they have chosen, and to resist using inferential statistics where inappropriate. In particular, we should be proactive in setting forth our choices of experimental design and our rationales for concluding that effects did or did not occur. This should be true of our manuscripts and, most especially, of our research grant proposals. Finally, we should be able to review manuscripts well enough to understand whether the statistics included are appropriate and appropriately described. Benefits to the field include more solidly based conclusions in the literature and perhaps greater respect for single-subject designs from colleagues who now dismiss them.

REFERENCES

Bordens, K. S., & Abbott, B. B. (1996). *Research design and methods: A process approach* (3rd ed.). Mountain View, CA: Mayfield.

Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance testing does and does not do. *The Behavior Analyst, 22,* 87–92.

Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *The Behavior Analyst, 21,* 111–123.

Harris, R. J. (Ed.). (1997). Special section: Ban the significance test? *Psychological Science, 8,* 1–20.

Hopkins, B. L., Cole, B. L., & Mason, T. L. (1998). A critique of the usefulness of inferential statistics in applied behavior analysis. *The Behavior Analyst, 21,* 125–137.

Kazdin, A. E. (1984). Statistical analyses for single-case experimental designs. In D. H. Barlow & M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavior change* (2nd ed., pp. 265–316). New York: Pergamon Press.

Krishef, C. H. (1991). *Fundamental approaches to single subject design and analysis.* Malabar, FL: Krieger.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way

we analyze data. *Current Directions in Psychological Science, 5,* 161–171.

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22,* 109–116.

Rosnow, R. L., & Rosenthal, R. (1996). *Beginning behavioral research: A conceptual primer* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* New York: Basic Books.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review, 57,* 193–216.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11,* 221–233.

The Behavior Analyst 1999, 22, 99–103 No. 2 (Fall)

Statistical Inference in Behavior Analysis: Having My Cake and Eating It?

Michael Davison
Auckland University

Using simple, nonparametric statistical procedures can formalize the process of letting data speak for themselves, and can eliminate the gratuitous dismissal of deviant data from subjects or conditions. These procedures can act as useful discriminative stimuli, both for behavior analysts and for those from other areas of psychology who occasionally sample our journals. I also argue that changes in publication policies must change if behavior analysts are to accurately discriminate between real, reliable effects (hits) and false alarms.

Key words: nonparametric statistics, Type I error, Type II error, conservatism

When I was invited to take part in the panel discussion that was the basis for this paper, I began to worry about my behavior in relation to statistics. I guess that one of the reasons that I was asked to take part was that I habitually use inferential statistics in my papers. Given that, perhaps, the unanalyzed life is not worth living, I started wondering why I did this. After searching my behavioral soul, my world-shattering conclusion is that my behavior is the product of my history. Let me lay bare some of this to you.

While I was working on my doctorate in Dunedin, New Zealand, my supervisor and I conducted some research concerned with what controlled the pause after reinforcement in fixed-ratio schedules. The data were exceedingly clear: The next ratio requirement was the major variable, though there were some effects of prior requirements. My supervisor suggested that I take these results to a conference in Australia, and I naturally agreed. "Oh, by the way, Australians are very keen on statistics, so you'd better do an analysis of variance," he said. So I did, being very keen to impress potential employers. I gave the paper, showed the data in all their glory, and then presented an ANOVA. Question time, and one person got up and said "That was the wrong ANOVA model, you should have used this one." A second person said the first person was wrong, it should have been that one. And then all hell let loose. There were no bouquets, questions about the data, the experimental procedure, nothing. There were no reinforcers. As we all know, one bad experience can sully your whole life, and this did. Until quite recently, I never did another parametric analysis of variance. I generalized from this to all parametric statistics, and have shied away almost completely from these henceforth.

Individual Differences

Over the many years that I have been writing for the *Journal of the Experimental Analysis of Behavior,* reviewers have vacillated from requiring no statistics, through requiring variance measures in data (but no inferential statistics), to requiring some sort of statistical treatment. From this extended training, and from my reading in the area, I came to the conclusion that not using statistics was a bad practice, and that at least some formalization of the process that we use to determine the meaning of data was required. One of the processes that upset me the most is the dismissal of results from single subjects or single condi-

I thank Susan Schneider and Douglas Elliffe for useful and informed comments on a draft of this paper.

Reprints may be obtained from Michael Davison, Department of Psychology, Auckland University, Private Bag 92019, Auckland, New Zealand (E-mail: m.davison@auckland.ac.nz).

tions. There is a great variation between researchers in the skill of arguing away deviant data, and I also suspect that a person's standing in the field can have a tremendous influence on whether such arguments are accepted in the editorial process or not. This last influence is perhaps reasonable, but for none, even the most respected, should it be allowed to wag the dog.

I understand the argument for dismissing deviant data. It is entirely possible that some subpopulation of my sample might, because of a different behavioral history, respond quite differently to an experimental manipulation. I probably would want to argue that this is a quantitative difference, rather than a qualitative difference. Within the parameters of a single experiment that did not investigate this quantitative difference, however, it would look qualitative. If this difference really exists, the behavior of one of my subjects might be different. But it would be hard to discriminate between "my finding is not general" and "I had one odd bird" because these are the selfsame conclusions. What I cannot conclude, however, is the all-too-common conclusion that the finding *is* general and that I had an odd bird. This conclusion is tantamount to making the assumption that the odd behavior comes from random error variance, rather than from systematic variation in an independent variable that has not been investigated. Knowing that behavior is multiply caused, I should conduct a follow-up experiment to determine the reason for the odd behavior. Arguing away the deviant subject, though, is functionally equivalent to using an increased N in an inferential statistical model. And as most behavior analysts recognize, inferential statistics gloss over deviant behavior and, depending on how large N is, will either accept or reject the null hypothesis (Hopkins, Cole, & Mason, 1998).

Sample Size

Just as dismissing deviant subject performance is not good practice, increasing the sample size in order to minimize individual differences is also not good practice. Sample size does matter, however. If it is the case that $N = 1$ is enough for a radical behaviorist, then, I am not (if this is a requirement for membership), and do not want to be, a radical behaviorist. I am interested in the external validity of my findings, and I feel the need to replicate across subjects—I need to have some idea that the results from a single subject were not an isolated, odd occurrence. But if the membership criteria include being able to show the same effect with all the subjects that we use, I will join. But the question is what N is enough? From my view, 3 or 4 is not enough, *especially* in those cases in which one of the subjects did something different, and its data are argued away. As you will see from my publications, I feel that I need 6 subjects to be able to get a good idea of what is going on. Given that most of my experiments are quite long, I can lose 1 subject and still feel happy. But, because I am interested in the behavior of individual organisms, using 6 subjects rather than 3 or 4 will increase the likelihood of my sampling the subject that behaves differently.

Nonparametric Statistics

I generally use nonparametric statistics when reporting my research. Such statistical procedures are appropriate for small numbers of subjects, but I cannot say, now, which is the chicken and which is the egg: Do I use 6 subjects because that gives me a result on a sign test if *all* subjects behave in the same way? Or do I use nonparametric statistics (rather than no statistics at all) because I generally have 6 subjects? I suspect that this is just a dynamic system that has come to feed on itself. However, I do believe that the use of such statistical procedures levels the playing field between researchers, and it does represent rather nicely the much more informal process of letting the data speak for themselves. Or would

do so, if most researchers used a similar number of subjects. I simply believe that 6 out of 6 (even 5 out of 5) is enough to give everyone the confidence that the findings have decent generality. I'd like to see this as a recommendation, rather than a rule.

Simple nonparametric statistics, I think, simulate the best behavior of researchers when they look at their data, and formalize the process of assessment. I think they can act as good discriminative stimuli for a reader's interpretation of the data, although there are dangers in doing so. If I simply report that a slope measure increased between two conditions for all 6 subjects, does this have the same impact as the additional test (significant on a sign test at $p < .05$)? In a sense it should, but for many readers, particularly those from other areas who occasionally sample behavioral journals, it is not. Behavior analysis needs such people to be impressed with its research and cannot afford to have them discard the work with the thought "I bet that isn't significant." Beyond statistics as a mathematical procedure, they are discriminative stimuli for behavior that because of the training of psychologists, have a substantial impact. Some behavior analysts may find this to be unfortunate, but if behavior analysis is to survive it has to live within a larger verbal community whose behavior is affected by the use of statistical tests.

One of the dangers of using statistics, however, is using them incorrectly. If you do the wrong test (like using the wrong ANOVA model!) you can get significance where there is none, and none where there is. Thus, especially for outside readers, behavior analysts also need to make sure they know what they are doing when they use statistics, and to report all germane parameters of the test (and the data used) so this can be checked. In this way behavior analysts end up being as much statisticians—perhaps even more so, given our bad press on this matter—than other psychologists.

	Effect found	No effect found
There is an effect	2 - P	
There is no effect		

	Effect found	No effect found
There is an effect		
There is no effect	2 - P	

Figure 1. Two published (P) findings of a "reliable" effect.

Conservatism

These considerations bring me to a further point: conservatism. Again, if joining the radical behavioral club requires me to be conservative in reporting differences, then I want in. Sidman (1960) was right when he insisted that our business is to look for generalities and invariances (Nevin, 1984) rather than differences. It seems to me that this is the real business of science, and that the statistical model (and even the nonstatistical, "let the data speak for themselves" model) can lead, through the interaction with publication processes, to psychology becoming a Type I error (Davison, 1998). The culture of psychology writ large produces students whose main aim in life is to find a significant difference (and hence, publish). We really do have to change this culture, and I believe it is much less prevalent in the experimental analysis of behavior than in the remainder of psychology.

As I have argued before, statistical analyses (or small N visual analyses) in combination with publication policies can have a disastrous effect on what is "known" (Davison, 1998). A signal-detection analysis, as shown in Figures 1 and 2, makes this clear. Take two situations, with an effect being reported in both. In Situation A (Figure 1), there is a real effect, and there are two published reports of this effect. The published data suggest a high

	Effect found	No effect found
There is an effect	2 - P	1 - U
There is no effect	0	0

	Effect found	No effect found
There is an effect	0	0
There is no effect	2 - P	8 - U

Figure 2. Reality. Two published (P) findings of a "reliable" effect, and some unpublished (U).

probability of a "hit," the reporting of a signal when one exists. In Situation B, there is no real effect, but there are also two published findings of an effect—two false alarms. Situation B, of course, could happen by chance. Subsequently, in both cases, no-effect results (misses in Situation A and correct rejections in Situation B) will be very hard to publish, and the journals report in both situations a 2:0 score line for "effect found." We cannot discriminate from the published data which effect is real. For both situations, the initial finding and its replication suggest an effect, but the real situation may be quite different. If misses in Situation A and correct rejections in Situation B were published, however, we might have more data to make the discrimination. For example, if there is one miss in Situation A (Figure 2), there is a reasonably certain decision ($p = .67$) that the effect exists. If in Situation B, however, there are eight correct rejections, then the probability that the effect exists is .2. Thus, the "reality" afforded by publication is strongly biased towards an effect being accepted regardless of whether it is real or not. It follows then, that unless we publish good-quality failures to replicate, we are systematically blinding ourselves to reality, and we cannot discriminate reality from fantasy.

It is also for reasons of conservatism, and the quest for generality, that I use nonparametric statistics. I would rather assume a lower level of measurement in my data than a higher level that my data may not reach. I would rather not assume that my data are normally distributed, because with $N = 6$, I cannot ever demonstrate that they are. I am aware that if the data were normally distributed, I could use more powerful tests, but I would rather deal with exact probabilities rather than approximations under a string of assumptions. I was strongly influenced by the first two chapters of Bradley (1968), which gives a very robust comparison of parametric and nonparametric tests. Finally, I suspect that our high levels of experimental control lead inevitably to the nonnormality of data distributions.

A wealth of nonparametric statistical tests are readily available. They range from binary comparisons, through analysis of variance with post hoc testing and orthogonal polynomial analysis, to regression, for all levels of measurement. Some of the tests, like rank randomization and normal-scores tests, are extremely powerful, with asymptotic relative efficiencies greater than 100% for nonnormal data in comparison with classical t and F tests. My particularly valued resources are Conover (1980), Ferguson (1965), and Marascuillo and McSweeney (1977), which provide an excellent coverage of useful tests. Although I recommend that readers look at other nonparametric tests, I try to keep my usage to simple sign and binomial tests, nonparametric trend tests, and Friedman analysis of variance. These seem to me to be honest and understandable and to fit closely with my own assessment of the importance of effects. Of course, unless I am well out of date, nonparametric tests do not offer analyses of interactions—for which I am eternally grateful. I have often argued, and will continue to do so, that significant interactions indicate wrong measures—yes, they can often be eliminated by transformations, but the goal of basic science must be to discover measures

that are independent of one another. Just think of the havoc that would be created if, in Ohm's law ($V = IR$), I and R had an interactive effect on V. (Actually, in any real system, this is exactly what happens as the amps that flow will warm the resistor and change its resistance, but I am talking fundamental, rather than applied, science here.)

Before concluding, though, with an exhortation that nonparametric statistics in our research should be used to provide a conservative playing field, there is one more problem that may arise. The subject matter of psychology is dynamically interacting behavior–environment systems (Davison, 1998). This may make experimental design difficult (independent variables are dependent, and dependent variables are independent, in some sense), and may make data analysis more difficult. In most of psychology and the experimental analysis of behavior, independent variables are nominal, or maybe ordinal. Either way, in any real system, they will have variance, and the arranged nominality or ordinality of the independent variables across subjects may be seriously compromised. It thus becomes important, even essential, to carry out analyses using procedures, such as linear regression (which are descriptive, rather than inferential) that take into account the continuous quantitative nature of the environmental input and its variance. Such regression procedures (known as structural relations and nonparametric regression procedures) are not widely used (see Davison & McCarthy, 1981, for an example) and are in need of further development and, especially, advertisement. It seems to me that assuming no variance in our "independent" variables is an error that is right up there with the global assumption that data are distributed normally (or that it matters not if they aren't)—nicely termed the "normal mystique" by Bradley (1968).

REFERENCES

Bradley, J. V. (1968). *Distribution-free statistical tests.* Englewood Cliffs, NJ: Prentice Hall.

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: Wiley.

Davison, M. (1998). Experimental design: Problems in understanding the dynamical behavior-environment system. *The Behavior Analyst, 21,* 219–240.

Davison, M., & McCarthy, D. (1981). Undermatching and structural relations. *Behaviour Analysis Letters, 1,* 67–72.

Ferguson, G. A. (1965). *Nonparametric trend analysis.* Montreal, Canada: McGill University Press.

Hopkins, B. L., Cole, B. L., & Mason, T. L. (1998). A critique of the usefulness of inferential statistics in applied behavior analysis. *The Behavior Analyst, 21,* 125–137.

Marascuillo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA: Brooks/Cole.

Nevin, J. A. (1984). Quantitative analysis. *Journal of the Experimental Analysis of Behavior, 42,* 421–434.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

The Behavior Analyst 1999, 22, 105–108 No. 2 (Fall)

Statistical Inference in Behavior Analysis: Useful Friend

John Crosbie
West Virginia University

Single-subject and statistical inference are virtually identical. With both techniques change is inferred when variability across conditions is sufficiently large to accommodate variability within conditions, replication is the final arbiter of whether change is likely to occur by chance, a large effect size is preferred to a small consistent difference, there are similar threats to internal validity, and generalizability of results is valued. Knowing how to use statistical inferential procedures would make behavior analysts more methodologically sophisticated. It would also help them to critically evaluate research in other areas of psychology, obtain research grants, and publish their research in diverse outlets, which would help others to see behavior-analytic work.

Behavior analysis has evolved a methodology that seems fundamentally different from that used in other areas of psychology. In behavior analysis the individual is the predominant unit of analysis, and inference of change is mainly visual. Hereafter in this article, this inferential method is labeled single-subject analysis. In most other areas of psychology, however, the group is the predominant unit of analysis, and inference of change is mainly statistical. Both of these methods have strong philosophical bases (Fisher, 1951; Sidman, 1960), and proponents of one position often treat proponents of the other position with suspicion and scorn. Because of this rivalry, some behavior analysts have concluded that *all* aspects of the groups approach, including statistical inference, are conceptually and methodologically flawed, or at least inappropriate for behavior analysis (e.g., Baer, 1977). Adopting such a position is like throwing out the baby with the bathwater. Behavior analysts have much to gain by studying inferential statistical procedures and using them when they are appropriate.

Preparation of this manuscript was supported by NIH Grant MH54195.

Correspondence should be addressed to John Crosbie, 4213 Colony Plaza, Newport Beach, California 92660 (E-mail: john.crosbie@worldnet.att.net).

Methodological Similarities

Consider the question that provided the impetus for all the articles in this series: "Should inferential statistical procedures be used in behavior analysis?" This question may be misleading. *All single-subject inference is statistical.* For example, compare how change is assessed with a simple single-subject design with one baseline and one treatment condition, and with a simple groups design with one experimental group and one control group. With the single-subject design, the baseline and treatment series are significantly different if there is minimal overlap of scores between the two conditions. As the variability of scores increases, a greater difference between the two series is required. With the groups design, the conditions are significantly different if a *t* test shows that the difference in means is sufficiently large to compensate for the variability within groups. As variability within groups increases, a greater mean difference is required to obtain statistical significance. The data-analytic logic of the two designs is virtually identical.

The two methods also have several other common methodological features: Replicability is the final arbiter of whether an effect is likely to occur by chance (Fisher, 1951; Keppel, 1982), effect size is more important than mere consistent difference (Co-

hen, 1977), history and maturation are important threats to internal validity (Cook & Campbell, 1979; Hersen & Barlow, 1976), and generalizability of effects is important (Baer, Wolf, & Risley, 1968; Keppel, 1982). When all of these methodological similarities are considered, it is clear that differences between the approaches are more apparent than real.

Why Behavior Analysts Should Use Statistical Inference

There are several reasons why behavior analysts should learn more about statistical inferential procedures, and use them when appropriate. The first, and most important, reason is that, in some situations, statistical inference is better than visual inference. With long, stable time series obtained under conditions of tight experimental control (e.g., many of those reported in the *Journal of the Experimental Analysis of Behavior*), visual inference is a fine procedure, and there probably is no need to use statistics. With short time series obtained without experimental control (e.g., many of those reported in the *Journal of Applied Behavior Analysis*), however, visual inference is not reliable, and cannot control Type I error (see Crosbie, 1993, for references to support both of these points). With such data, some inferential supplement is required to achieve the behavior-analytic goal of only accepting large, reliable, robust effects. To illustrate, here are two situations in which statistical analysis was useful with single-subject data.

In one study (Crosbie & Kelly, 1994), college students completed several sets of programmed material (Holland & Skinner, 1961) in a multiple schedule. Sets were randomly assigned to experimental conditions. Because sets differed in terms of number of frames and difficulty, the percentage of frames answered correctly differed markedly between sets, and that variability obscured the analysis. For each subject the experimental condition had better results than the control condition, but not for each comparison. Consequently, it was difficult to determine with visual inference whether the experimental and control conditions produced significantly different results. Statistical inference, however, was able to accommodate the inherent variability in the data. If there were no difference between conditions, then scores from the experimental condition would be greater than scores from the control condition on approximately half of the comparisons. There may be local variations in the departure from .5 probability, but they would be only minor. If, however, there was a difference between conditions, then experimental scores would be greater than control scores with a probability that is significantly greater than .5. Even though scores came from the same subjects, random assignment of sets to conditions overcame any potential problem of order effect or autocorrelation (Edgington, 1982). Across 4 subjects in that study, there were 40 experimental sessions and 40 control sessions, so there were 40 comparisons for the sign test (Siegel, 1956). Experimental scores were greater than control scores on 30 of the 40 comparisons, which has a probability of .0011. Although inherent variability in the data made it difficult to *see* clear difference between conditions for individual subjects, the sign test showed that the conditions were different.

In another study (McClannahan, McGee, MacDuff, & Krantz, 1990), the authors assessed the effects of providing feedback to group-home parents concerning the personal appearance of their children at school. Feedback was provided on a multiple baseline across homes. There were 21 observations per home, 6 to 11 observations during baseline, and the data were variable with an increasing trend. Those features made visual inference difficult. In Home 1, there seemed to be an increase during the feedback condition, but the increase was slight. In Home 2, it was not clear whether there was an

increase during the feedback condition, or whether the baseline trend merely continued during the feedback condition. Visual inference was inconclusive with such data. A customized interrupted time-series analysis procedure (ITSACORR; Crosbie, 1993, 1995), however, was able to accommodate the variability, short phases, and autocorrelation to show that there was a significant increase in Home 1 but not in Home 2.

Although several authors have argued that statistical analysis is too lenient for behavior analysis (e.g., Baer, 1977), my research shows that ITSACORR is much more conservative than is visual inspection. Thus, the statistical test is used as an additional hurdle, which ensures that only strong effects survive.

There are other reasons why behavior analysts need to know how to use statistical inferential procedures. Many people understand the rules of their native language much better after they have studied another language. That also is true with inferential procedures. By studying another form of inference, behavior analysts will learn more about single-subject analysis. Only then will they understand the logic of all inference, and move beyond uncritical acceptance of one particular position. Furthermore, there are methodological issues such as autocorrelation and controlling Type I error that are important for both approaches, but can be studied most easily with statistical procedures such as Monte Carlo simulation. Even if behavior analysts never use statistical inferential procedures, knowing how to do so will improve their methodological sophistication, and therefore make them better behavior analysts, psychologists, and scientists.

Being able to use statistical inferential procedures also will help behavior analysts to obtain research grants, and let them conduct their research as they want. Many members of study sections are unfamiliar with, and even hostile towards, single-subject designs. If grant proposals are described purely in single-subject terms, the probability of funding is reduced. Furthermore, even if the grant is funded, the reviewers may ask for groups designs and analyses that are not what the researcher wants, nor what would be best for the research. If researchers are unable to explain clearly and persuasively why the proposed single-subject design and analysis accommodates all the reviewers' concerns and is most appropriate, they are defenseless. With funding agencies, knowledge is power.

Being able to understand statistical inferential procedures also permits behavior analysts to read and critically evaluate research published in diverse journals, and thereby stand on the shoulders of giants in other areas. Furthermore, behavior-analytic research could be published in diverse places, which would help to spread knowledge about environmental effects on behavior.

Conclusion

Before I am branded a heretic, let me stress that I do not advocate a fundamental change in how behavior analysis is conducted or conceptualized. What I recommend is that behavior analysts should add statistical inferential procedures to their toolbox, because that tool is useful for scientific, educational, political, and evangelical purposes. For too long scientists who use formal statistical procedures and behavior analysts have treated each other with suspicion and scorn, to the detriment of both groups and to science in general. It is time for détente.

REFERENCES

Baer, D. M. (1977). "Perhaps it would be better not to know everything." *Journal of Applied Behavior Analysis, 10,* 167–172.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1,* 91–97.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-*

experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61,* 966–974.

Crosbie, J. (1995). Interrupted time-series analysis with short series: Why it is problematic; how it can be improved. In J. M. Gottman (Ed.), *The analysis of change* (pp. 361–395). Mahwah, NJ: Erlbaum.

Crosbie, J., & Kelly, G. (1994). Effects of imposed postfeedback delays in programmed instruction. *Journal of Applied Behavior Analysis, 27,* 483–491.

Edgington, E. S. (1982). Non-parametric tests for single-subject multiple schedule experiments. *Behavioral Assessment, 4,* 83–91.

Fisher, R. A. (1951). *The design of experiments* (6th ed.). Edinburgh: Oliver & Boyd.

Hersen, M., & Barlow, D. H. (1976). *Single case experimental designs: Strategies for studying behavior change in the individual.* New York: Pergamon.

Holland, J. G., & Skinner, B. F. (1961). *The analysis of behavior.* New York: McGraw-Hill.

Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

McClannahan, L. E., McGee, G. G., MacDuff, G. S., & Krantz, P. J. (1990). Assessing and improving child care. *Journal of Applied Behavior Analysis, 23,* 469–482.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* New York: Basic Books.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* Tokyo: McGraw-Hill Kogakusha.

The Behavior Analyst 1999, 22, 109–116 No. 2 (Fall)

Statistical Inference in Behavior Analysis: Experimental Control is Better

Michael Perone
West Virginia University

Statistical inference promises automatic, objective, reliable assessments of data, independent of the skills or biases of the investigator, whereas the single-subject methods favored by behavior analysts often are said to rely too much on the investigator's subjective impressions, particularly in the visual analysis of data. In fact, conventional statistical methods are difficult to apply correctly, even by experts, and the underlying logic of null-hypothesis testing has drawn criticism since its inception. By comparison, single-subject methods foster direct, continuous interaction between investigator and subject and development of strong forms of experimental control that obviate the need for statistical inference. Treatment effects are demonstrated in experimental designs that incorporate replication within and between subjects, and the visual analysis of data is adequate when integrated into such designs. Thus, single-subject methods are ideal for shaping—and maintaining—the kind of experimental practices that will ensure the continued success of behavior analysis.

Science is a social enterprise, and the standards of scientific evidence are established by consensus. From this perspective, the objective of research design and data analysis is straightforward: to convince an audience of skeptical colleagues that a particular interpretation or inference is justified. The rules of statistical inference, set forth in classic texts and promulgated in mandatory graduate courses, provide an agreed-upon solution. By following these rules—and rejecting the null hypothesis with a p value of less than .05—investigators assure their peers, and themselves, of the significance of their findings. Statistics guide investigators to inferences about their data that can be expressed in objective, quantitative terms. Indeed, the inferences seem to arise automatically from the application of the statistical formula, as implied by the term most commonly used to describe the process: *statistical* inference. When scientific inferences are produced by a formula, the *investigator* is relieved of a burdensome responsibility, and science itself is protected from the frailties of human judgment, which is error-prone and subject to an assortment of troubling biases.

Requests for reprints should be sent to Michael Perone at the Department of Psychology, West Virginia University, P.O. Box 6040, Morgantown, West Virginia 26506-6040 (E-mail: mperone@wvu.edu).

Limitations of Statistical Inference

Or so it seems. Unfortunately, statistics offer the investigator no panacea, and no self-respecting statistician would claim otherwise. Despite the central role played by null-hypothesis statistical tests throughout the biological, behavioral, and social sciences, fundamental problems have been recognized for some time (e.g., Bakan, 1966; Lykken, 1968; Meehl, 1967). By 1970, the criticisms of statistical inference had drawn enough attention from psychologists to warrant a book provocatively entitled *The Significance Test Controversy* (Morrison & Henkel, 1970). But actual use of statistical analysis has not changed much since then; null-hypothesis testing is as robust as ever. Cautions may be decreed by textbook authors and professors in statistics courses, but when students and investigators are confronted with real research problems, they are beguiled by the reassuring directness of statistical procedures, which offer simple rules for answering a host of practical questions ("How many subjects per cell?"). In return, textbooks and professors seem more than willing to

offer simple recipes for cooking up the answers ("Run Cohen's power analysis program and see what it says"). Browse through an assortment of statistics texts, and you will find many with handy tables and flow-charts to guide the reader to just the right test for the data at hand, putting the task of analyzing the results of an experiment on the same level as looking up a telephone number. The appearance of "point-and-click" software for statistical analysis has made matters worse. When asked by a puzzled associate editor to explain an unusual statistic in a manuscript submitted for publication, more than one author has responded by providing the name and version number of the software package.

The bottom line is this: Too much research design and data analysis is performed without thinking. Thompson (1998) voiced his objection this way:

> Most researchers mindlessly test [the null hypothesis] because most statistical packages only test such hypotheses. This . . . does not require researchers to thoughtfully extrapolate expected results from the previous literature or from theory. Instead, science becomes an automated, blind search for mindless tabular asterisks using thoughtless hypotheses. (p. 799)

Although statistical analysis has its defenders (e.g., Dixon, 1998; Hagen, 1997, 1998; Wilcox, 1998), the criticisms of years past continue to cause trouble, and debate about statistical strengths and weaknesses is being repeated and expanded by a new generation of psychologists and statisticians (e.g., Cohen, 1994; McGrath, 1998; Tryon, 1998). The greatest strength of statistical inference—the automatic, objective, reliable assessment of data, all independent of the skills or biases of the investigator—is a mirage. Research summarized by Tryon indicates that statistical tests are routinely misinterpreted by investigators publishing in our best journals, and even by statisticians themselves. "How much more susceptible to misinterpretation," he asks, "are the vast majority of other less well quantitatively trained psychologists?" (p. 796).

On the Search for Methodological Imperatives

Eventually some investigators discover that there is no good cookbook for delicious servings of research design and data analysis. But too many still see design and analysis as obstacles to good research rather than an integral part of it. If they have grants, they hire statistical consultants. If they are graduate students, they make sure a statistics professor is a member of their dissertation committee. The prevailing attitude is that the framing of research questions can proceed apart from the methods employed to answer them.

The attraction to formulas and rules is not confined to investigators who favor group-statistical approaches. Professors who teach courses in single-subject research design are confronted by students of behavior analysis seeking, for example, rules about the criteria used to decide that behavior has reached a steady state. Over the years many students have reported that they adopted the criterion recommended by Sidman (1960) in his classic *Tactics of Scientific Research.* But Sidman never offered such a recommendation. In answer to the question "How does one select a steady-state criterion?" he explained, "There is . . . no rule to follow, for the criterion will depend upon the phenomenon being investigated and upon the level of experimental control that can be maintained" (p. 258). On what basis, then, is one to decide? Sidman pointed to the investigator's "accumulated experience and good experimental judgment" developed in the course of "designing and carrying out steady-state experiments" (p. 261).

We are left with a dilemma: Group-statistical methods incorporate tidy sets of rules, but the rules lead to less than satisfactory results, even in the hands of veterans. Single-subject methods

seem to offer no rules at all. What guidance, then, is to be offered the student embarking on a research career?

The answer may be found in three critical notions in the passages quoted from Sidman's (1960) book: experience, experimental control, and judgment. These are recurring themes in Sidman's treatment of research tactics and, indeed, throughout the historical development of behavior analysis. Understanding the role they can play in scientific research is the key to appreciating why statistical inference has not been and need not become a major factor in the experimental analysis of behavior.

Experience

Behavior analysts' interest in single-subject as opposed to group-statistical research may be regarded as the result of an inductive process arising from intense interactions with data and various practical considerations, rather than deductions from a well-developed philosophy of science. The sophisticated philosophical justification for single-subject research came later.

Behavior-analytic methods, of course, derive from the work of Skinner, whose graduate training antedated the widespread adoption of the group-statistical approach made possible by Fisher (1925). Skinner's early research involved single-subject designs; most of the experiments reported in his seminal work, *The Behavior of Organisms* (Skinner, 1938), used only 4 rats. But as large-group methods gained favor within psychology in the late 1930s, Skinner, then an assistant professor at the University of Minnesota, gave them a try. He and Heron built a set of 24 operant chambers and cumulative recorders, interconnected so that the recorders displayed mean performances for the entire group of 24 rats, as well as subgroups of 12 and 6. Skinner said that he and Heron "thus provided for the design of experiments according to the principles of R. A. Fisher, which then were coming into vogue" (Skinner, 1956/1972, p. 113). Skinner was enthusiastic about the approach; he reported that "the possibility of using large groups of animals greatly improves upon (our) method ... since tests of significance are provided for and properties of behavior not apparent in single cases may be more easily detected" (Skinner, 1956/1972, p. 113). But Skinner's enthusiasm soon faded:

> In actual practice that is not what happened. . . You cannot easily make a change in the conditions of an experiment when twenty-four apparatuses have to be altered. Any gain in rigor is more than matched by a loss in flexibility. We were forced to confine ourselves to processes which could be studied with the baselines already developed in earlier work. We could not move on to the discovery of other processes or even to a more refined analysis of those we were working with. No matter how significant might be the relations we actually demonstrated, our statistical Leviathan had swum aground. (Skinner, 1956/1972, pp. 113–114)

Skinner, the consummate tinkerer, was quite willing to scout about for new ways to conduct experiments. He rejected group-statistical methods not because they collided with his radical behaviorist epistemology, but rather because his experience revealed that they insulated the investigator from the behavior of the subject. The ongoing interaction between experimenter and data that had characterized his earlier work—and led to his innovations in apparatus, measurement, and theory—could not be sustained in group-statistical research. Skinner returned to the experimental analysis of individual behavior, and directed his energies to developing stronger methods of experimental control that would obviate the need for statistical inference.

Experimental Control

The tension between group-statistical and single-subject methods is created by the relative roles played by experimental control in the two approaches. For Skinner and other advocates of single-subject research, group-statistical methods are ill suited to the development of strong forms of experimental control over behavior, in

part because the group methods are unwieldy and in part because the nature of statistical analysis reduces the investigator's motivation to establish such control. The sensitivity of a statistical test is a direct function of the number of subjects, and weak control can be tolerated if the number is large enough. Averaging data across many subjects can hide a multitude of sins: The experimental treatment may fail to affect the behavior of some subjects, and may even lead to contrary effects in others. As a consequence, statistically significant results based on large sample sizes are not persuasive. Given a sufficiently large sample, statistical significance is assured. Meehl (1967) pointed out that the only question is whether the direction of the statistical difference will support the investigator's hypothesis. Under these circumstances, the probability of support is a lofty .5—hardly a rigorous experimental challenge.

In single-subject research, by comparison, treatment effects are clarified not by increasing statistical sensitivity but rather by improvements in experimental control. Individual differences are not averaged into obscurity as statistical error, but instead are regarded as revealing the limits of the control being exercised.

As a case in point, consider a situation encountered in the course of an experiment on "observing behavior" in adult humans (Perone & Baron, 1980). The main response was pulling a plunger mounted underneath a table. On the table was a console with colored stimulus lamps and several response keys. In the critical conditions, pressing the "observing" keys on the console would turn on colored lights correlated with the schedules of monetary reinforcement associated with the plunger response. During preliminary training, 1 subject adopted an unusual response topography: He tied one end of his bootlace to the plunger and the other end to the leg of his chair, put his feet on the table, and executed the response by rocking back and forth. When a monetary reinforcer was presented (an occasional event given the intermittent nature of the schedule), the subject repositioned himself and pressed a button on the console required to collect the reinforcer, then resumed the rocking motion. This topography was wholly compatible with the monetary schedule, which involved only the plunger response, but the investigators worried that it would interfere with the acquisition of the observing response, because the observing keys would usually be out of the subject's reach. To block the chair-rocking topography, the investigators replaced the chair with a wheeled stool. The subject reacted by sitting on the floor, tying his bootlace to the plunger and pulling the other end, and occasionally standing up to collect reinforcers. The new topography was no better than the old one. Finally, the investigators placed a limited hold on the collection button: Once a monetary reinforcer was earned, the subject had just 1 s to get up and collect it before it was canceled. This contingency was effective in moving the subject onto the stool in front the console, with the collection button and the observing keys within easy reach. When the critical phase of the experiment finally commenced, the subject acquired the observing response and his data fell in line with those of the other subjects.

The close interaction between investigator and subject fostered by the single-subject approach allowed a potential disaster to be identified and averted. The troublesome individual difference was not relegated to a statistical error term, but was eliminated by suitable adjustment in the experimental procedure. What would have happened in a group experiment? Perhaps the absence of a conditioned reinforcement effect in the problem subject would have been overlooked, if it did not appreciably affect the group mean. Or, if detected, the negative result might have been attributed to the regrettable but inevitable appearance in the sample of a recalcitrant subject whose person-

ality leads to sabotaging experimental goals. Of course, nothing about group designs prevents the kind of corrective action taken in this case. But the ready acceptance of individual differences and other forms of "error variance" seems more amenable to the theory and practice of group-statistical research than to single-subject research.

As noted elsewhere (Perone, 1991), a successful experimental science is one that exerts high degrees of control over its subject matter. The ability to control variables that affect behavior is prerequisite to the study of steady states. Thus, because single-subject designs require investigators to seek strict levels of control, their adoption encourages the development of an experimental science of behavior.

Judgment

Although some discussions of group-statistical methods may suggest otherwise, human judgment is an unavoidable component of the scientific enterprise. Investigators must exercise their best judgment repeatedly over the course of a research project. At the outset they must decide what line of investigation is likely to make a contribution to knowledge. Then they must devise appropriate experimental designs and procedures, often balancing competing interests based on convenience, economy, and the availability of apparatus and personnel. They must puzzle over the measures to employ, analyses to conduct, which results are worth reporting, and the implications of the results for contemporary theoretical debate. They must decide how methods, results, and arguments should be conveyed to the scientific community in the form of grant applications, publications, and professional presentations. Sometimes they must decide whether a negative outcome should spur a reappraisal of one's experimental strategy or abandonment of a cherished theoretical position. All these judgments and more are a matter of routine for active scientists regardless of their discipline, theoretical predilections, or epistemological convictions.

The adoption of group-statistical methods does not eliminate the need for an investigator's sound judgment, nor does the adoption of single-subject methods guarantee it. The two kinds of methods do, however, place different judgmental burdens on the investigator. And because of the relative rarity of single-subject methods, the burdens of that tradition are often misunderstood. Perhaps the greatest misunderstanding revolves around the so-called "visual analysis" of data.

When it comes to analyzing experimental results, the difference between group-statistical and single-subject methods is sometimes characterized along these lines: In group research, inferences about causal relations between independent and dependent variables are guided by precise, sophisticated statistical tests free of subjectivity and bias. In single-subject research, investigators stumble along with only a simple graph of the results to inspect unaided, leaving their causal inferences susceptible to all manner of idiosyncratic influences. Again, the absence of codified rules for conducting the visual analysis is seen as the culprit. Kazdin (1982) expressed the problem this way:

> Perhaps the major issue pertains to the lack of concrete decision rules for determining whether a particular demonstration shows or fails to show a reliable effect. The process of visual inspection would seem to permit, if not actively encourage, subjectivity and inconsistency in the evaluation of intervention effects. (p. 239)

Research is available to bolster this criticism. Investigators given session-by-session graphs of concocted behavioral data and asked to judge the presence of treatment effects may disagree with one another, be swayed by seemingly minor details of the graphic presentation, overlook small but reliable effects, or see effects when they are absent (DeProspero & Cohen, 1979; Knapp, 1983; Matyas & Greenwood, 1990; but see Parsonson & Baer, 1992, for a more appreciative account of visual analysis).

The rejoinder is that criticism of visual analysis is based on a profound misunderstanding. Indeed, the very term *visual analysis*—and the research into it—does not adequately represent the process as it occurs in actual research. Perhaps the problem can be traced to the comparison with statistical analysis. Statistical tests are conducted after an experiment is completed and the results are in. At that point, the investigator is left to sift though the data and seek evidence that an effect was brought about by the experimental manipulations. Critics of visual analysis seem to believe that it is merely an unsophisticated version of the same process: After the experiment the investigator draws a graph of the results and decides about the influence of the independent variable. But in practice no single-subject experiment is conducted in such a fashion. Visual analysis is an ongoing activity throughout the experiment; indeed, it is an integral part of the experimental analysis and as such it cannot be separated from the methods employed to collect the data in the graphs.

The point may be clarified by restating it with a more appropriate emphasis: *Experimental* analysis is an integral part of visual analysis. By this account, it is a mistake to suggest that investigators in the single-subject tradition prefer the visual inspection of graphs over statistical analysis. What *is* preferred is an experimental analysis so thorough, so powerful in its control over the subject matter of interest, that cause–effect relations are plain to see. The experiment may be regarded as any other scientific instrument, such as a microscope, whose resolution is painstakingly refined until the object of study comes into clear focus. The behavior analyst does not rely on unaided senses to see causal relations in behavior any more than the biologist relies on the naked eye to see subcellular objects. The adequacy of visual analysis depends on, and can be no greater than, the adequacy of the instrument aiding the investigator's vision, and in the study of behavior the instrument is the experiment. To be valid, a single-subject experiment must show that behavioral states can be replicated at will in different subjects and at different times within the same subject. Replication thus establishes the investigator's success in identifying and controlling relevant variables and confirms the adequacy of the stability criteria that guide the investigator's decisions about the attainment of steady states (see Baron & Perone, 1998, and Perone, 1991, for detailed discussion of the validity of single-subject experiments). In this connection, it is noteworthy that the previously cited research questioning the adequacy of visual analysis does not address the role of replication across subjects, nor does it express doubt about the conclusions of actual single-subject research.

Conclusions

The question prompting this essay is the role inferential statistics should play in behavior analysis. Ever since group-statistical methods gained favor in psychology, behavior analysis has drawn criticism for its devotion to single-subject methods. This essay has tried to show that the criticisms are based on an exalted and erroneous view of the power of statistical inference, one that regards statistical tests as a set of tried and true rules that reliably and inevitably guide investigators to objective answers for their experimental questions. In practice, however, statistical inference is not so simple. The rules, such as they are, have proven difficult to apply, even in the hands of statisticians, and the underlying logic of null-hypothesis testing has drawn fire since its popularization by Fisher nearly 75 years ago. Paradoxically, the criticism most often leveled against single-subject methods—that they do not ensure consistent outcomes across investigators—seems to apply equally to group-statistical methods.

Tests of statistical inference may

have their place in psychology, and perhaps even in behavior analysis. But there is no room for the unthinking methodological orthodoxy that often accompanies statistical inference. Perhaps the trouble started when Campbell and Stanley (1963) proclaimed that the only "true experiment" is one with random assignment of subjects to treatment groups. Campbell and Stanley directed their monograph to field researchers in education, and it seems unlikely that they intended to dismiss single-subject experiments (or, for that matter, virtually all natural science before 1925) as invalid. But by parroting Campbell and Stanley's monograph with insufficient thought or circumspection, several generations of textbooks on psychological research methods have surely had that unfortunate effect.

Whatever methods are adopted by behavior analysts, let us ask that they be adopted thoughtfully. The cookbook recipes sometimes associated with statistical inference are easy to criticize, but more thoughtful statistical applications may be welcome. In the same vein, it must be recognized that the demand for cookbooks is not altogether absent from the behavior-analytic community. Sidman (1960), as he wrote his *Tactics,* was perhaps the first to feel the demand. His response was to steadfastly refuse to offer any recipes. Instead, he asked his readers to think analytically about their research questions, to explore new procedures, and to learn from experience—in short, to develop good experimental judgment.

The present view, derived from the insights and advice offered by Sidman and Skinner, is that in a science of behavior good judgment is shaped by intensive interplay between investigator and subject in the course of experimental analysis. Group-statistical methods seem ill suited to the task, tending to insulate the investigator from the immediate results of experimental operations and reducing the motivation for seeking and exercising strong forms of control. By comparison, single-subject methods put investigator and subject into repeated contact, and force the investigator to identify and control variables relevant to the object of study. Thus, the methods are ideal for shaping—and maintaining—the kind of experimental practices that will ensure the continued success of behavior analysis.

REFERENCES

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Baron, A., & Perone, M. (1998). Experimental design and analysis in the laboratory study of human operant behavior. In K. A. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 45–91). New York: Plenum.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 573–579.

Dixon, P. (1998). Why scientists value p values. *Psychonomic Bulletin & Review, 5,* 390–396.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver & Boyd.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Hagen, R. L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist, 53,* 801–803.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* New York: Oxford University Press.

Knapp, T. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5,* 155–164.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23,* 341–351.

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist, 53,* 796–797.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115. (Reprinted in D. E. Morrison & R. E. Henkel, Eds.,

The significance test controversy, pp. 252–266. Chicago: Aldine, 1970)

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy.* Chicago: Aldine.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Hillsdale, NJ: Erlbaum.

Perone, M. (1991). Experimental design in the analysis of free-operant behavior. In I. H. Iversen & K. A. Lattal (Eds.), *Techniques in the behavioral and neural sciences: Vol. 6. Experimental analysis of behavior, Part 1* (pp. 135–171). Amsterdam: Elsevier.

Perone, M., & Baron, A. (1980). Reinforcement of human observing behavior by a stimulus correlated with extinction or increased effort. *Journal of the Experimental Analysis of Behavior, 34,* 239–261.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Skinner, B. F. (1938). *The behavior of organisms.* New York: Appleton-Century.

Skinner, B. F. (1972). A case history in scientific method. In B. F. Skinner (Ed.), *Cumulative record* (3rd ed., pp. 101–124). New York: Appleton-Century-Crofts. (Original work published 1956)

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53,* 799–800.

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist, 53,* 796.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.

The Behavior Analyst 1999, 22, 117–121 No. 2 (Fall)

Statistical Inference in Behavior Analysis: Discussant's Remarks

Richard L. Shull
University of North Carolina–Greensboro

A collection of essays on the roles of inferential statistics in behavior-analytic research prompted consideration of five issues: (a) the acceptance of research that focuses on the behavior of individual organisms; (b) the need to apply methods thoughtfully; (c) the heuristic value of statistical description; (d) the treatment of aberrant data in the search for general principles; and (e) the role of derived measures in the search for invariances.

It was a pleasure to read these thoughtfully prepared essays on the possible roles of inferential statistics in behavior-analytic research. All the papers in the set contain insightful comments and useful analyses, and there is just enough diversity in outlook to encourage reflection. I will restrict my comments to five themes.

1. It was hardly surprising that none of the authors advocated a wholesale endorsement of group-based research designs and the methods of inferential statistics to validate research findings. There seemed to be good agreement that the goal of most research carried out by behavior analysts is the formulation of cause–effect laws that apply to the behavior of individual organisms and that apply generally. At one time, long ago, it might have been possible to wonder seriously if behavioral phenomena were too disorderly to permit cause–effect relations to be demonstrated in the behavior of individual subjects. No one can seriously wonder about that now. One of the great contributions of behavior-analytic research since the 1930s has been to provide overwhelming evidence that orderly relations, of considerable generality and involving complex behavioral phenomena, can be demonstrated in the behavior of individual subjects (Sidman, 1960). That point is now widely understood and acknowledged. Research conducted in the behavior-analytic style is now commonly reported in many of the most respected journals identified with mainstream psychology and other fields. Behavior analysts can feel justly proud that some of the most reliable, general, and practically important functional relations found in behavioral science have been generated from the behavior of individual organisms through behavior-analytic research methods. Thoughtful people outside the field of behavior analysis know this. There should be no need for defensiveness on this score.

2. So why, then, is the frequency of inferential statistics in behavior-analytic research increasing (Baron, 1999)? And is this increased frequency cause for concern? Several of the contributors seem clearly worried that the quality of behavior-analytic research may be declining and that reliance on the methodology of inferential statistics is contributing to this decline. Doing science well is difficult. It requires discipline, complex discriminations, intense concentration on the issue of interest, enormous patience, a dose of compulsive-like perfectionism, substantial persistence in the face of failure (but knowing also when it is wise to switch), a large capacity for work, a knowledge base that is broad and deep, and a certain willingness to take risks. A theme that runs through several of the papers (Ator, 1999, Branch, 1999,

Correspondence concerning this paper can be addressed to Richard L. Shull, Department of Psychology, Box 26164, UNCG, Greensboro, North Carolina 27402-6164 (E-mail: rlshul@uncg.edu).

and Perone, 1999, especially; see also Verplanck, 1998) is that researchers may latch on to the methods of inferential statistics and thereby avoid the hard work and tough choices that good science requires. Branch (1999), for example, notes that relying on the results of a statistical test can be a way of avoiding responsibility for the soundness and validity of the results that one is reporting. Ator (1999) gives us a "top-10" list of reasons for relying on statistics to salvage poor-quality data instead of carrying out the additional work that could result in a genuinely substantial contribution. Her list exemplifies what may be a growing, broader concern with trends in how research is conducted, particularly under the contingencies of an academic environment (e.g., Harzem, 1990; Thomson, 1994; Vicente, 1998). Put bluntly, most of the reasons that she lists are excuses for intellectual laziness. Perone (1999) stresses the importance of being in close contact with the experimental subject matter and working constantly to improve experimental control; he notes how reliance on inferential statistics can weaken those important components of scientific behavior by seeming to provide a shortcut to success (albeit an illusory shortcut).

I find these concerns to be serious, valid, and effectively expressed. At the same time, I was glad to see acknowledgment that these unfortunate effects are not *necessary* effects of using inferential statistics. Most of the professional statisticians with whom I have talked would abhor the kinds of sloppiness and laziness noted by Ator (1999), Branch (1999), and Perone (1999). Fortunately, a fair number of researchers who carry out group-based research designs demonstrate great concern for such things as improving experimental control and generating such large and consistent effects that there can be little doubt that the cause–effect relations are relevant to the behavior of individual organisms. (I confess, though, to finding it far more convincing and satisfying to see the relevance to individual organisms demonstrated directly instead of indirectly by inference from group-based data.) My sense is that all the authors would regard it as very foolish to ignore an interesting, well-conceived, and well-carried-out study simply because it was conducted under a group-based design. Inferential statistics and group-based designs may enable certain kinds of unfortunate practices, but they don't force such practices. Moreover, conducting research in the behavior-analytic style is hardly a guarantee of high-quality research.

Indeed, it may be possible to forget that methodology is a means to the end of discovering important cause–effect principles that are relevant to the behavior of individual organisms. It is not itself the end. As Platt (1964, p. 351) so vividly cautioned us:

> Beware of the man of one method or one instrument, either experimental or theoretical. He tends to become method-oriented rather than problem-oriented. The method-oriented man is shackled; the problem-oriented man is at least reaching freely toward what is most important.

The major concern, then, seemed to be not so much about the use of inferential statistics but about their use in an unthinking, formalistic way for the wrong reasons; for example, what Branch (1999, p. 89) called the "slavish adherence to significance testing." I think all the authors would endorse Ator's (1999, p. 96) plea for "thoughtfulness in experimental design" and Perone's (1999, p. 115) hope that "Whatever methods are adopted by behavior analysts, let us ask that they be adopted thoughtfully."

3. Two of the contributors (Crosbie, 1999, and Davison, 1999) see a heuristic role for inferential statistics. Approached intelligently, the use of inferential statistics can highlight and clarify some of the factors that influence the judgments and decisions that experienced researchers make. One need not agree fully with Crosbie (1999, p. 105) that "the data-analytic logic of the two designs is virtually identical" to appreciate his point that some as-

pects of inferential statistics can usefully be conceptualized as analogous to some aspects of the decision making by behavior analysts unaided by formal statistical analyses. As discussed by all the authors (and with particular thoroughness by Perone, 1999), the decision making by scientists—for example, deciding when a baseline is sufficiently stable to introduce the next condition or deciding when an effect has been demonstrated reliably—involves many complex discriminations and consideration of matters such as what already is known about related phenomena. The experienced, successful scientist is likely to have acquired many of the prerequisite skills through direct experience and may be in a poor position to identify for others the bases of his or her astute decisions. It would be a great benefit if we could formulate guidelines for effective decision making (e.g., about stability). That could make the training of effective scientists more efficient (Skinner, 1969), or, in Davison's words (1999), create a more "level playing field" (p. 100).

This goal is certainly laudable. A danger, however, is that guidelines can be formulated that do not really match the complexity and subtlety of the required discriminations and judgments (Skinner, 1969). Following such guidelines can result in what Ator (1999, p. 95) aptly described as "rule-governed behavior run amok." And that is the problem with following unthoughtfully the rules of inferential statistics (or of following analogous rules for determining stability; Perone, 1999): Those rules fail to capture the richness, subtlety, and complexity of the conditions that actually guide the decision making of effective, experienced scientists. What is needed is an experimental analysis of such decision making. As just one small example, we might ask what stimulus features of data plots actually lead experienced behavior analysts to continue or stop a condition. The relevant stimulus features are surely molar and relational, and it may be that statistics could provide a useful language for describing such features. Once those features are characterized, we might be able to use statistical variables to help teach the relevant discriminations to students—the outcome of some statistical analysis being, in effect, a second opinion supporting or challenging the student's judgment. Killeen's (1978) insightful treatment of stability criteria exemplifies this approach (see also Baer & Parsonson, 1981; Parsonson & Baer, 1992). I understood the positive message of Crosbie's (1999) and Davison's (1999) papers as encouraging the empirical analysis of scientific decision making so as to create more effective guidelines that can be passed on to individuals entering the field.

4. Davison (1999) is justifiably concerned about the practice of explaining away the data from aberrant subjects. If our aim is to formulate cause–effect laws that are general, then it will not do to be satisfied with showing a relation that holds, under a specified set of conditions, only some of the time or for only some of the experimental subjects. Of course, if data were reported as values averaged over groups of subjects, then we could not know, without additional information, whether a reported average trend was characteristic of the effect for most of the individuals in the group, for some of the individuals, or for none of the individuals. And the fact that an average trend might be statistically significant does not help much. Focusing on the individual-level trends reveals the exceptions and can (or should) encourage experimental analysis to identify the factors responsible for the differences (Branch, 1999; Perone, 1999; Sidman, 1960).

Davison's (1999) important concern is reminiscent of Johnson's (1932, pp. 306–307) earlier lament regarding imprecise assertions about the generality of findings:

> To appraise their assertion, we should first ask what they mean by the little word "general." It has, in fact, two principal meanings, which we dare not interchange in the same discourse. In

the language of mathematics and of exact science, it means "subject to no exceptions." In the language of social conversation, and of vague, listless, and slovenly description, it means "subject to exceptions." For example, consider the assertion, "Red-haired women, in general, are high-tempered." In the language of exact description, the assertion means that red-haired women, without exception, are high-tempered. In the language of slovenly description, it means that all red-haired women are high-tempered except those who are not. It is impossible to make an assertion more vague than that. Its extent is indeterminate; one cannot tell whether it includes all members of the subject-class, some members, or none.... Certainly, any equation is "general" in the slovenly sense of that term, for it will describe all experimental results satisfactorily, except for those which it will not.

5. Several of the authors (e.g., Baron, 1999, Branch, 1999, and Perone, 1999) comment on the problems that can arise when the reported effects are based on average values. Average values of performance may bear little resemblance to any of the performances contributing to the average. Cause–effect principles based on such averages can, therefore, be quite misleading about what will occur when a real individual organism is exposed to the relevant causal conditions. This problem can occur whether the average is taken over different organisms or over different performances of the same organism. These concerns are important and valid, but there is a balancing consideration as well. Our goal as scientists, as Davison (1999) reminds us, usually is to discover invariances: relations that hold true regardless of the specific circumstances of the particular experimental arrangement. Sometimes principles based on average values or on other derived descriptive measures will be more general (i.e., less contingent on particular conditions) than principles that are based on the descriptive characteristics of the more immediately apparent performances. Again and again in science, what is most immediately apparent turns out not to be what is most general and important for effective understanding. To take a specific example, a smoothly increasing gradient of responding is seldom observed under fixed-interval (FI) schedules when the focus is on the performance records from individual interreinforcement intervals; the smooth FI scallop emerges from averaging (cf. Baron & Leinenweber, 1994). Yet it is possible that the averaged gradient expresses a more general, fundamentally important characteristic of temporal control than does the individual-interval-level description. The more common pause-run performance pattern might be telling us more about the effects of particular features of our experimental preparation (e.g., the use of a relatively effortless response, the placement of a highly salient stimulus at the response location, etc.) than about general characteristics of temporal control. As Baron and Leinenweber (1994, p. 17) suggested,

> Whatever the limitations of the average scallop as a depiction of response patterns, averaging procedures are unavoidable if one is to consider operant performances as response probabilities. When the present results are viewed in this light, the irregular variations in the break-and-run and single response patterns in the cumulative records may be seen to reflect orderly tendencies for responses to become increasingly likely as time passes (i.e., a scalloped pattern of responding).

Again, a message that I believe all the authors endorse is that we should avoid slavish adherence to rules and instead take a thoughtful approach to all these sorts of questions.

REFERENCES

Ator, N. A. (1999). Statistical inference in behavior analysis: Environmental determinants? *The Behavior Analyst, 22,* 93–97.

Baer, D. M., & Parsonson, B. S. (1981). Applied changes from steady state: Still a problem in the visual analysis of data. In C. M. Bradshaw, E. Szabadi, & C. F. Lowe (Eds.), *Quantification of steady-state operant behaviour* (pp. 273–285). Amsterdam: Elsevier.

Baron, A. (1999). Statistical inference in behavior analysis: Friend or foe? *The Behavior Analyst, 22,* 83–85.

Baron, A., & Leinenweber, A. (1994). Molecular and molar analyses of fixed-interval performance. *Journal of the Experimental Analysis of Behavior, 61,* 11–18.

Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance

testing does and does not do. *The Behavior Analyst, 22,* 87–92.

Crosbie, J. (1999). Statistical inference in behavior analysis: Useful friend. *The Behavior Analyst, 22,* 105–108.

Davison, M. (1999). Statistical inference in behavior analysis: Having my cake and eating it? *The Behavior Analyst, 22,* 99–103.

Harzem, P. (1990). Seven bad reasons for doing research and one good one. *Journal of Veterinary Medical Education, 17,* 28–29.

Johnson, H. M. (1932). Some follies of "emancipated" psychology. *Psychological Review, 39,* 293–323.

Killeen, P. R. (1978). Stability criteria. *Journal of the Experimental Analysis of Behavior, 29,* 17–25.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Hillsdale, NJ: Erlbaum.

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22,* 109–116.

Platt, J. R. (1964). Strong inference. *Science, 146,* 347–353.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Skinner, B. F. (1969). An operant analysis of problem solving. In B. F. Skinner, *Contingencies of reinforcement* (pp. 133–171). New York: Appleton-Century-Crofts.

Thomson, K. S. (1994). Scientific publishing: An embarrassment of riches. *American Scientist, 82,* 508–511.

Verplanck, W. S. (1998). Statistical inference: Why wheels spin. *Behavioral and Brain Sciences, 21,* 223–224.

Vicente, K. J. (1998). Four reasons why the science of psychology is still in trouble. *Behavioral and Brain Sciences, 21,* 224–225.