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 (Sid-

man, 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 rele-

vance 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 sel-

dom 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.