METHODOLOGICAL AND ASSESSMENT CONSIDERATIONS IN APPLIED SETTINGS: REVIEWERS' COMMENTS

The large number of nonexperimental manuscripts submitted to JABA and the lack of well-defined criteria for evaluating them, has necessitated formulation of a separate editorial policy (see policy statement on page 404 of this issue). The manuscript by Alan E. Kazdin was submitted before the policy concerning nonexperimental manuscripts went into effect and was reviewed by three established researchers who have made substantial methodological contributions to the field (Montrose M. Wolf, Murray Sidman, and L. Keith Miller were Reviewers A, B, and C, respectively). On the basis of the reviewers' comments, it was decided to publish the manuscript with only minor changes. However, because of the fundamental importance of many of the issues discussed in Kazdin's paper and because of the lack of clear agreement on some of these issues, as exemplified in the reviewers' comments, these comments are presented below.

COMMENTS BY REVIEWER A

It was a pleasure to read this paper on experimental design by Alan Kazdin. The author makes many important points that JABA readers and researchers need to consider. He makes these points simply and eloquently. Kazdin's suggestion that behavior modification researchers need to consider the advantages of control group design is, I think, particularly well made and important. Also, he will do many readers a service by introducing them to the experimental design considerations of Campbell and Stanley (1963). I do have two suggestions, however, for the author to consider:

1. The paper focuses on what the author describes as "reinforcement" programs as if the only variable used in these programs was reinforcement. Then, the author describes how most behavior modification research does not attempt to analyze the role of any particular variable but instead usually involves a "package" of treatment variables such as instructions, training, practice, and reinforcement. As a result, as Kazdin points out, it is not usually possible to determine the amount of the effect that is attributable to any one of these variables. Kazdin attacks this strategy as a serious weakness in behavior modification research. It seems to me that the author himself is confounding two very important but different strategies, that of *basic-theoretical research* and of *applied research*.

In basic-theoretical research, the goal is to conduct a demonstration experiment that will support a *principle*, like the principle of reinforcement or the principle of extinction, or the principle of stimulus generalization, *etc.* This is unlike technological research, which is directed at finding a solution to a particular applied problem. In technological research, there is often no attempt to isolate the contribution of any of the variables that are applied simultaneously toward the solution of the problem. In technological research, the primary goal is to find some way to modify specific behaviors. Thus, in technological research rather than simply introducing one variable at a time and studying its effect, the investigators usually introduce multi-variable packages to see whether the whole thing has any effect. If indeed the package does have an effect and *then* if there is some *practical* or *theoretical* reason for examining the role of the individual components these roles can then be teased out. Frequently, the role of the individual components is trivial in terms of implications for either practice or theory. For example, if we find that a teacher going to a child and praising him and at the same time patting him on the shoulder increases that child's rate of doing his arithmetic, do we really care whether it was the praise statement or the pat? As long as the teacher must be next to the child anyway in order to deliver either of the consequences it may be just as easy to

deliver both. The same is true for most complex behavior modification treatment packages. They may be composed of many classes of variables and each of these classes of variables may in turn be composed of a great many components. As Kazdin pointed out, in many studies it is not clear whether it is the instruction, the training, or the token reinforcement system that produces the effect. This is true. On the other hand, each of those classes of variables is made up of many, many variables. Even if it were discovered that the token system played a major role in a change, one would still not know what specific variables in the token reinforcement component were functional. It could be the color of the token, its texture, some of the things that back it up (but probably not all), *etc.* In other words, to isolate the role of every conceivable variable in a complex treatment package would take a researcher literally forever. It is only in the laboratory that we have the luxury of reasonably *pure* variables, and often even this is illusory.

Rather than encouraging behavior modification researchers to spend more time isolating the role of each of the several variables in their treatment packages, it seems to me that they should instead be helped to consider when such component analyses would have practical or theoretical significance. For example, if an element in a treatment package is expensive or difficult to teach or to monitor, then there would be a definite practical reason for analyzing its role. Or if an element in a treatment package is novel for behavior theory, then it would be worthwhile teasing out its role so that its implications for behavior theory would be clear.

It should be pointed out even more forcefully that researchers frequently do not indicate that their treatment packages are multi-dimensional. As indeed Kazdin has pointed out, they are usually made up of more than reinforcement, and researchers often do not explicitly list the other variables like the instructions, announcements, modelling, and practice that may go into

the treatment program as well as the reinforcement. Kazdin could have done us a wealth of good by prodding researchers to become more explicit in describing all of the variables of the treatment programs, thus making the descriptions of these treatment programs in JABA more likely to facilitate replication.

2. The author stated that response frequency as a dependent variable is "the most useful" in behavior modification research. However, Kazdin did not describe how he determined that this measure was "the most useful". Thus, the statement is ambiguous and should have been deleted, since there is no way to evaluate it. As a matter of fact, the Bijou *et al.* (1968) article that the author cites to support his statement does not represent an example of the use of frequency data. While Bijou sometimes uses the term "frequency", one can see by reading his procedure section that he does not use frequency as a response measure at all. Rather, he samples during each interval of a few seconds length whether or not the behavior occurs or does not occur. Thus, a response with an extremely low frequency, such as one per session, but with a long duration, such as the session length, would be noted to occur within every one of the intervals and thus would represent 100% on Bijou's graphs (which are all percentage graphs). None of Bijou's graphs display frequency.

As for the articles by Ferster (1953) and Honig (1966), which are also used to support the author's point about response frequency as "the most useful" measure, they are both animal research articles. The relevance of this animal research for contemporary applied research escapes this reviewer. The fact is, I think that frequency as a dependent variable had a great deal of use in the animal laboratory but has not generalized well to applied research. A cursory review of the response measures used in JABA articles indicates that frequency is one of the least popular dependent variables. The Bijou occurrence sampling procedure, time sampling, amount (such as pieces of litter), and per cent correct (such as in academic work) seem to be

more favored than frequency as dependent variables by applied behavior analysis researchers.

Kazdin might have helped researchers and students by categorizing all of the different dependent variables that are currently used. On the other hand, the statement that frequency is "the most useful" dependent variable may mislead the casual reader.

Recommendation. This paper is potentially very important to students and JABA readers who may not be familiar with the numerous critical points made by Kazdin. I strongly recommend publication and look forward to re-reading it myself soon in JABA.

COMMENTS BY REVIEWER B

This paper is a worthwhile contribution, and I recommend publication. The author might, however, consider the following points before publication:

1. Methods for assessing within-subject effects are not limited simply to reversals and multiple baselines. For example, one may also introduce quantitative variations in the variable of interest (the question of whether the complete elimination of the variable, as in the "reversal design", is simply an extreme instance of quantitative variation is not a simple problem); one may use the finding to predict other effects; the credibility of a finding can be increased by specification of the conditions under which it does *not* occur; and irreversibility may be *explainable* as a *legitimate* outcome of an experimental manipulation. Perhaps more thought should be given to these matters.

2. The problem of response and stimulus specification—as these differ between subject and experimenter—is more general than is indicated here. It is a problem in any experiment, regardless of the formal design, and it may be misleading to single out the multiple-baseline design as a specific instance.

3. Cumulative effects over time, interactions among individuals (social effects), and progressive baseline changes are all real problems in

carrying out studies within a non-laboratory setting, as in a token-economy project. These effects are not a function of any particular "design". They are often unavoidable, and may either limit the generality of any study or even preclude certain kinds of investigations. Yet they do not necessarily invalidate all such studies. The question is at the nub of the distinction between laboratory and clinical research, and is responsible for the different standards in evaluating these two kinds of research. The mistake is to judge experimental and clinical research by the same criteria (although they are certainly related). I would like to have seen a more positive contribution along these lines in this paper, and not simply the critical and perhaps mistaken slant that is actually expressed.

4. With respect to internal and external validity, it would have been useful to suggest some answers to the problems the author poses. Many of the points are good, but many investigators who are already aware of them would appreciate some positive suggestions, and this would be a more meaningful contribution. The use of control groups can sometimes help, but only at the most primitive level of analysis.

COMMENTS BY REVIEWER C

I found this paper very well written, scholarly, and interesting. It presents an overview of our present research design methodology—particularly the strengths and weaknesses of the reversal and multiple baseline designs. It raises questions about some ways in which incorrect conclusions can be reached (under the rubric of internal and external validity). And it suggests that we should use both control groups and measures of supplementary behaviors more frequently. These are all extremely important considerations.

In spite of my considerable sympathy for the paper, I find it extremely difficult to review. Perhaps the major reason is that no very clear criteria have been established, or even suggested,

for evaluating methodological papers. In view of this, it is essential that the reviewers explicitly state their criteria. The only unique criteria that I have been able to come up with is: "Does the manuscript specify a procedure or tool that might be useful in conducting behavioral research?" Thus, I shall review the manuscript from this point of view, by paraphrasing what the author seems to be saying in the major parts of the manuscript and providing an evaluation of each part.

PART I: Within-subject designs have important advantages over between-group designs.

These advantages are briefly described and include: (a) intersubject variability is bypassed; (b) the behavior of the individual subject is determined; (c) lawful effects of variables can be discovered; (d) individual differences are more readily observed.

Evaluation. These and other advantages are more clearly and more extensively discussed in a variety of other sources (e.g., Baer, Wolf, and Risley, 1968; Bernard, 1953; Sidman, 1960).

PART II: The reversal design assumes that the experimental effect is reversible.

This is clearly an important consideration to keep in mind when evaluating the feasibility of using a reversal design in a particular setting. The author adds the further assertion that "Indeed, we would hope that they are not always reversible because this is tantamount to demonstrating only slight resistance to extinction." And, further, that behavior modification changes must be "transient" in order to use this design.

Evaluation. The general point is clearly functional in planning an experiment. One must make a guess as to whether the behavior will reverse before blindly choosing a reversal design. In my opinion, most readers of JABA already understand this point and attempt to use it in designing their experiments. The author might have made a useful contribution by suggesting guidelines for making this guess. He might have sifted through the literature, or his

own experience, to select some clues for reversibility that could be translated into rules of thumb. He might have suggested some techniques whereby the chances of reversibility occurring are enhanced (the reinforcement of an alternative, incompatible response, for example). I feel that development of his theme along these lines might well have served a useful function for readers of JABA. However, his present treatment is too general and it stopped short of providing functional guides to action. It therefore does not meet my criteria of functionality.

I would like to add that his assertion that reversal designs can be used only to study transient effects is misleading. Any effect of consequences is transient once the consequences are removed unless there are other consequences to take their place. As I understand the goals of applied behavior analysis, it is to learn how to modify an environment to produce specified changes in behavior. Once the original environment is reinstated, we must ordinarily expect the behavior to revert to its original baseline rate. It is only if the original environment had a consequence that was too weak to initiate a behavioral change but that is strong enough to maintain such a response (once initiated) that the behavior should not be expected to revert (Baer and Wolf, 1970). While we may frequently work in environments with such weak consequences, we also frequently seek to discover how to engineer changes in an environment so that the changed environment will produce specified changes in some target behavior. Such an engineered environment need be in no sense transient. I would guess that Kazdin is also interested in such engineered environments; however, his phrasing could easily be misinterpreted.

PART III: The multiple-baseline design cannot be used where the dependent variables (responses, subjects, or settings) are correlated.

This is clearly an important consideration to be examined before selecting a multiple-baseline

design. If an increase in one behavior will in some way increase the other behaviors of a multiple-baseline design, then no independent effect of the treatment can be established.

Evaluation. Again, this general point is functional in planning an experiment. However, it is hard for me to imagine a JABA reader who does not understand this. If there is one, he is probably neither a user nor a consumer of multiple-baseline designs.

In this case also, the author could have attempted to provide some rules of thumb for guessing whether several behaviors will be correlated. His advice on this point is not very functional, however. The author suggests that whether the dependent variables are correlated or not must be learned from experience, although in some (unspecified) instances "the investigator can rely on the well-documented experience of others". I am not clear whether the author's review of such correlations was meant to be a guide or simply a demonstration that the problem does arise. I would like to see an attempt to review known cases with the goal of providing some guidelines for predicting situations in which this problem might arise. For example, generalization is probably one source of correlation between variables. Therefore, a possible rule of thumb would be to select behaviors that are topographically as different as possible and that occur in stimulus situations that are as different as possible.

PART IV: Different experimental designs may suggest different experimental conclusions for the same questions.

Kazdin gives two examples in which the same variables are manipulated in a between-group design and a within-subject design, and in which the results suggested different conclusions for each set of data.

Evaluation. Frankly, the author already outlined a number of reasons why the results of between-group designs are questionable (e.g., "lawful effects of variables may be obscured"). I agree with criticisms of group designs and

feel they raise questions about the significance of this part of the manuscript. Therefore, I would first have to ask if this same type of interaction between the type of design and the results obtained from each type also applies to single-subject designs. If such an interaction seemed plausible, then I would be interested in some rules of action. How can we identify potential interactions and avoid them?

PART V: An experimental result may be caused by an unspecified component of the experimental manipulation.

Kazdin gives three examples of this problem. First, instructions to subjects are often given simultaneously with a change in contingencies. Thus, either the new contingency, the instructions, or both acting together may be the cause of any change. "In several studies, the effect of instructions . . . cannot be separated from the reinforcement contingencies themselves." Second, when some event, such as praise or tokens, is given contingent upon a response, any change might also have been caused by the "informational feedback" implicit in the delivery of the contingent event. Thus, we cannot be sure that a simple feedback system might not be just as effective as our more complicated token systems. Third, the agent who implements any experimental manipulations may change his behavior as he changes the conditions. Thus, we could not be sure that the manipulations themselves are responsible for the change—it could be the agent's behavior that is responsible for the change in the subject's behavior.

Evaluation. I cannot argue with Kazdin's points. However, they are less relevant to *applied* research than they are to *basic* research. Many authors of research articles have taken to carefully qualifying their conclusions by talking about the "package" of events that constitute the treatment. Thus, the token treatment is often taken to consist of delivering a token that can be exchanged for goodies, providing feedback about the agent's target behavior implicit in that token delivery, the social reinforcement

that is necessarily involved in most token systems, and perhaps even the unspecified changes in teacher behavior induced by the token delivery (such as added attention, Mandelker, Brigham, and Bushell, 1970). By limiting their conclusions to "packages", applied researchers can present technologies that will produce a desired effect without necessarily being able to pinpoint all of the effective variables.

It would be a functional message to suggest to researchers that they limit their firm conclusions to the entire package that they used in the treatment. Another alternative is, of course, what might be called "component analysis". That is, the experimental analysis of the components of the treatment package may be separately analyzed. While Kazdin gives several examples of such analyses, his examples are more directed to pointing out the possibility of unspecified variables being responsible for experimental results. Perhaps these examples could be reworked to point out the possibility of such "component analyses". I would hope for some guidance as to when such an analysis might be important for an applied researcher. For example, when a particular component of a procedure is hard to implement or expensive to use, there may be applied reasons for evaluating the contribution of that component to the total treatment package.

PART VI: The particular experimental procedure that is selected may produce results that should not be generalized.

Kazdin gives several examples of this: first, the efficacy of shock punishment depends upon the manner in which the shock is introduced. That is, a high initial shock level will tend to eliminate responding indefinitely, while the introduction of such a level following the gradual increase of shock level may not totally eliminate the response. Second, each time a reinforcement period is followed by an extinction period, the rapidity of extinction will increase. And third, many experimental results may depend upon an observer being present in the situation and

cannot be generalized to situations in which there is no observer.

Evaluation. I have difficulty finding any functional advice to draw out of this observation. Any thorough researcher, such as Azrin in his series of punishment experiments (see Azrin and Holz, 1966), might have a good chance to discover the adaptation to shock punishment as well as the effect of gradual *versus* abrupt change in the level of the shock. But what advice could be given to someone: "be thorough"? We usually hope that we are. Could that rule be specified in behavioral terms so that it translates into a procedure or tool? I cannot get to it. The second example is somewhat similar. It deals with the stable state in a multiple schedule situation. In that state, extinction will occur more quickly until some stable state is reached appropriate for the schedule of reinforcement (Ferster and Skinner, 1957). Again, what do we suggest to someone who might like to learn something from the example?

And I guess I do not agree too much with the "observer effect". If the effect depends on the observer, then by all means alter the environment by inserting a permanent observer (or if that is too costly, look for alternative treatments). But that is perfectly generalizable as a piece of behavioral engineering. Frankly, I doubt that most applied behavior analysts would agree with the assertion that "the influence of observers is assumed not to affect the differential performance of . . . school children . . .". Measuring both teacher and student behavior is the basis for any alterations that anyone can make in a classroom. How can the student be expected to behave differently if contingencies are not placed on his behavior? And, likewise, how can teachers be expected to act differently in applying those contingencies to the children, if their behavior is not observed and consequated?

PART VII: Control groups are sometimes useful.

Clearly, the use of control group sometimes provides information necessary to the practical

use of a behavior modification method. If you wish to increase the rate of studying and find that some treatment will do just that, you still might want to find out if that has produced a concomitant increase in an achievement test. To evaluate the importance of the gain in the achievement test implies comparison with an untreated or "normal" group. The author gives a number of such examples.

Evaluation. Again, I cannot take exception to the general point. But I would like to see some rules or guidelines for identifying situations in which a control group should be used. It does not help me much to know that they are sometimes useful.

However, perhaps an even more important point might be to provide some guidance on how to avoid using control groups. Virtually all research on the effects of contingency management procedures in college classrooms has relied on between-groups designs. What a shame that our knowledge of the effects of such procedures is based on such a weak design. Fortunately, single-subject analyses of college teaching procedures are finally appearing (*e.g.*, Mawhinney, Bostow, Laws, Blumenfeld, and Hopkins, 1971). And even the analysis of changes in achievement test scores has now been approached through single-subject analysis (Miller and Weaver, 1972).

PART VIII: Socially important behaviors that might be modified as an incidental effect of some treatment should be measured.

Kazdin suggests several interesting samples of such a strategy. For example, the severity of inappropriate responses may decrease concomitantly with the treatment to reduce frequency of inappropriate responses. Children rewarded for appropriate behaviors also showed changes in IQ scores and arithmetic achievement levels (although I cannot figure out why Kazdin refers to these as "cognitive" variables when they are samples of the child's behavioral repertoire, and I do not know how "anxiety" was measured in this same example).

Evaluation. I thoroughly agree with Kazdin's suggestion. Supplementary behavioral observations would be a great help in understanding the effects of our environmental modifications. Why should we not know if children laugh (appropriately) as much when tokens are used as when anarchy reigns supreme? Many potential consumers of our technology wonder. Why not tell them?

Again, some guidelines for how to select such supplementary behaviors might be helpful. Perhaps asking socially relevant "judges" what they consider to be important in a situation would provide important clues to selecting such supplementary measures. The work of Wolf and his associates with "social validity" is a major step in the direction (*e.g.*, Phillips, Wolf, and Fixsen, *in press*). It may provide the most important guideline in this regard.

* * * * *

Now, how do I evaluate the overall manuscript? As I see it, the manuscript as it is constructed at present fails the test that I proposed. For the most part, it does not specify procedures and tools that might be useful in conducting behavioral research. In particular, it does not provide rules for predicting when a behavior will not be reversible, when several behaviors will be intercorrelated, when to conclude that a "package" of elements is responsible for an effect or how and when to do a "component analysis" of that package, how to be thorough, when to use control groups and how to avoid their use, or how to select supplementary variables for measurement.

It should be clear that I find many of Kazdin's observations interesting and thought-provoking. There may be merit in provoking thought in JABA readers. But what are the criteria for when something is sufficiently interesting or the provoked thought worthwhile? I personally cannot formulate them. I am afraid that to get into the business of provoking thoughts at all would open a Pandora's box for JABA's editorial board. At present, we are primarily in the

business of publishing sound behavioral research and secondarily in the business of publishing manuscripts that offer potentially useful research tools. This is a very narrow perspective. Let's keep it that way.

I therefore recommend that JABA not publish this very fine manuscript.

REFERENCES

- Azrin, N. H. and Holz, W. C. Punishment. In W. K. Honig (Ed.), *Operant behavior: areas of research and application*. New York: Appleton-Century-Crofts, 1966. Pp. 380-477.
- Baer, D. M. and Wolf, M. M. The entry in natural communities of reinforcement. In R. Ulrich, T. Stachnik, and J. Mabry (Eds.), *Control of human behavior, Volume 2*. Glenview, Illinois: Scott, Foresman and Company, 1970. Pp. 319-324.
- Baer, D. M., Wolf, M. M., and Risley, T. R. Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1968, 1, 91-97.
- Bernard, C. *Introduction to experimental medicine*. New York: Grolier Publishing Company, 1953.
- Campbell, D. T. and Stanley, J. C. Experimental and quasi-experimental designs for research and teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963. Pp. 171-246.
- Ferster, C. B. The use of the free operant in the analysis of behavior. *Psychological Bulletin*, 1953, 50, 264-274.
- Ferster, C. B. and Skinner, B. F. *Schedules of reinforcement*. New York: Appleton-Century-Crofts, 1957.
- Honig, W. K. *Operant behavior: areas of research and application*. New York: Appleton-Century-Crofts, 1966.
- Mandelker, A. V., Brigham, T. A., and Bushell, D. The effects of token procedures on a teacher's social contacts with her students. *Journal of Applied Behavior Analysis*, 1970, 3, 169-174.
- Mawhinney, V. T., Bostow, D. E., Laws, D. R., Blumenfield, G. J., and Hopkins, B. L. A comparison of students' studying-behavior produced by daily, weekly, and three-week testing schedules. *Journal of Applied Behavior Analysis*, 1971, 4, 257-264.
- Miller, L. K. and Weaver, F. H. A multiple baseline achievement test. In G. Semb (Ed.), *Behavior analysis and education—1972*. Lawrence, Kansas: Support and Development Center for Follow Through, Department of Human Development, University of Kansas, 1972. Pp. 393-399.
- Phillips, E. L., Wolf, M. M., and Fixsen, D. L. Achievement Place: development of the elected manager system for administering routine tasks. *Journal of Applied Behavior Analysis*, (in press).
- Sidman, M. *Tactics of scientific research*. Chicago: Basic Books, 1950.