Reaction:

Program Evaluation: The Independent and Dependent Variables

Lee Sechrest

University of Michigan

Clearly Oetting's paper is filled with good advice, and there are many points on which I might have commented. Although couched in fairly general terms, Oetting's paper could be read with profit by almost any counseling (or other applied) psychologist. However, his paper did key off two concerns of mine about which I would like to comment in some detail. Program evaluation can be boiled down to three issues: What are the independent variables? What are the dependent variables? And how are the two linked? Generally speaking, in treatises on evaluation, especially with respect to what is called "design" of evaluations, more attention is paid to the problem of linking independent and dependent variables than to the characteristics of either, especially those of independent variables. I wish, in my comments, to make some observations about independent variables and dependent variables.

Independent Variables

Independent variables are, of course, those that define the intervention or treatment. Since in most instances it is reasonably evident what the treatment is, relatively little attention is paid to the independent variable(s). That relative lack of concern is a mistake, for it affects two of the four important aspects of validity of an experiment: construct validity and external validity (Cook & Campbell, 1979). Construct validity has to do with our ability to define a treatment accurately. It is essential in interpreting the results of an evaluation study that we know exactly what the intervention or treatment was. That we know what a treatment was or is should not be taken for granted. External validity has to do with the legitimacy with which we may generalize about our findings in relation to other contexts. If we do not know precisely what intervention we evaluated, we can hardly be on safe grounds in generalizing about its usefulness in other settings.

The problem of construct validity begins with the question whether we actually know what the nature of the treatment is that is being tested. The fact that there are so often handy labels for certain types of interventions should not mislead us into thinking that there is equivalence between the label and the treatment. Several years ago Kazdin and Wilcoxin (1976) showed that a variety of nonspecific treatment effects may obscure the true nature of therapeutic interventions. Effects such as those often labeled placebo and expectancy may be mistaken for those one would like to attribute to more definitive and specific treatments. In any given study, the labeled treatment variable may be confounded with therapist enthusiasm (or lack thereof; not everything nonspecific may be therapeutic). Entering into counseling may, in a particular setting, be associated with a variety of other changes in persons' lives that may have effects mistakenly attributed to counseling specifically.

In fact, however, treatments are rarely assessed with sufficient care to make it possible to describe them with great accuracy. They are even less often described in sufficient detail that anyone else could expect to replicate them. Just to provide a simple example, consider counselor personality and interpersonal style. Nearly anyone would agree that counselor personality is likely to be critical to counseling outcome, but rarely, if ever, are the personalities of counselors carefully assessed in evaluation studies, and I, for one, have never seen a published study in which counselor personalitites were described unless the study was specifically on the importance of personality variables. In an analysis of programs designed to rehabilitate delinquents, Redner and I (Sechrest & Redner, 1979) found that in only six of 29 studies was the treatment variable sufficiently well described that one could understand it at all, let alone have any hope of replicating it. Clearly, part of the problem lies in the pressure for space in our journals, and detailed descriptions of interventions probably seem like low priority items. In my opinion, those descriptions are vital and should be required. If they cannot be accommodated in the journals, editors should at least require that they be submitted for examination (inadequacy of documentation should be a basis for rejection) and that authors agree to make them available upon request.

Beyond providing descriptions of interventions, we should also work toward development of quantitative indices of treatment *strength*. I and my colleagues have written extensively about this issue in other places (Sechrest & Redner, 1979; Sechrest, West, Phillips, Redner & Yeaton, 1979; Sechrest & Yeaton, 1981a; Yeaton & Sechrest, 1981a), and it does not require much elaboration here. Suffice it to say that until we know more about the strength (in a sense akin to dose) at which our treatments are being delivered, we can make only inadequate judgments about their effects. If counseling has little impact on juvenile delinquents, is that because counseling is inherently ineffective or because the counseling actually tried is weak (e.g., inadequately trained counselors, poor rationale for counseling, too few sessions, and so on)? I am persuaded that a great deal of what we attempt to evaluate would, upon careful examination, have had little chance of producing the effects we look for.

We also need to assure ourselves of treatment *integrity*, which has also been discussed in the previously listed publications by me and my associates. Treatment integrity refers to the degree to which treatments are actually carried out as planned. Even the best conceived counseling program can be expected to fail if it is not properly implemented. Proper implementation cannot be taken for granted; it must, in fact, be documented. Were counselors actually as well trained as expected? Were treatment protocols carefully followed? Did clients actually attend sessions for which they

were scheduled? Negative answers to these and many other critical questions would have to be given for many programs that have been evaluated and found wanting. No evaluation of a medical procedure would be accepted if it could not be established that the procedure was carried out in strict accordance with a specific protocol. We should not settle for less in evaluating human service delivery programs such as counseling.

Dependent Variables

Oetting mentions some of the problems involved in identifying and measuring relevant and meaningful dependent measures, and he notes in passing that there may be a discrepancy between a statistically significant outcome and one that has any practical importance. In fact, we have only a very poor understanding of the problem of assessing the magnitude of any effect we produce. Yeaton and I (Sechrest & Yeaton, 1980, 1981a, 1981b, 1981c; Yeaton & Sechrest, 1981a, 1981b) have been preoccupied for several years with ways of estimating effect sizes. Again, we need not go over that ground in detail here, but I do note that purely statistical estimates such as are provided in measures of variance accounted for are not likely to be of much help. Yeaton and I believe that we need empirical evidence for effect size estimates. The most useful evidence is likely to come in the form of either normative data or behavioral implications.

By normative data, we mean that when closely comparable outcome measures have been used in a number of different studies, it should be possible to determine the differences between experimental and comparison groups within each study and then to aggregate them across studies to provide a "norm" by which new treatments or evaluations could be judged. We might at least then be in a position to say of a treatment that it produced an effect larger than 75% of other treatments, or that since it produced an effect exceeded by 90% of other treatments, it would not appear promising.

Many of our outcomes are not measured in a metric that has immediate and inherent meaning. For example, so what if the anxiety scale scores of a group of treated clients are reduced by an average of 3.8? What is to be made of a 1.7 mean scale point improvement on the evaluation factor of a semantic differential? Yeaton and I believe that more work needs to be done to demonstrate the behavioral implications of our outcome measures. In what way would someone with a lower anxiety scale score be different? Kazdin (1977) has made a somewhat similar suggestion, calling for "social validation" of behavioral changes produced in treatment. If we knew, for example, that an anxiety score of 14 was associated with consistently better peer ratings on personal competence than an anxiety score of 18, we would have some idea of the value of reducing anxiety scores by that much. Until we have demonstrated what difference our differences make, we cannot expect them to be accorded much respect.

I will conclude by reiterating my respect for the wide range of Oetting's paper. Obviously it cannot be taken as a literal manual for program evaluation, but it is useful in giving an overview of the field and the problems and issues that reside in it. If it stimulates readers to think, to search further, and to give more consideration to the necessity for evaluation, it will have served its purpose. The paper provides no grounds for either pessimism or sloth.

References

Cook, T. D., & Campbell, D. T. *Quasi-experimentation: Design and analysis for field settings.* Chicago: Rand McNally, 1979.

Kazdin, A. E. Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, 1977, *1*, 427-452.

Kazdin, A. E., & Wilcoxin, L. A. Systematic desensitization and nonspecific treatment effects: A methodological evaluation. *Psychological Bulletin*, 1976, *83*, 729-758.

Sechrest, L., & Redner, R. Strength and integrity of treatments in evaluation studies. *Criminal Justice Research Reports*. Washington, DC: National Criminal Justice Reference Service, National Institute of Justice, 1979.

Sechrest, L., West, S. G., Phillips, M., Redner, R., & Yeaton, W. H. Strength and integrity of treatment: Some neglected problems in evaluation research. In L. Sechrest, et al. (Eds.), *Evaluation Studies Review Annual* (Vol. 4). Beverly Hills: Sage Publications, 1979.

Sechrest, L., & Yeaton, W. H. Limitations on the interpretation of measures of effect size in evaluative research. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 1980, 382-385.

Sechrest, L., & Yeaton, W. H. Assessing the effectiveness of social programs: Methodological and conceptual issues. *New Directions for Program Evaluation*, No. 9. San Francisco: Jossey-Bass, 1981, 41-56. (a)

Sechrest, L., & Yeaton, W. H. Empirical bases for estimating effect size. In R. F. Boruch, P. M. Wortman, & D. S. Cordray (Eds.), *Reanalyzing program evaluations*. San Francisco: Jossey-Bass, 1981, 212-224. (b)

Sechrest, L., & Yeaton, W. H. Estimating magnitudes of experimental effects. *Journal Supplements Abstract Service, Catalog of Selected Documents in Psychology*, Nov. 1981, Doc. No. 2355. (c)

Yeaton, W. H., & Sechrest, L. Critical dimensions in choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, 1981, *49*, 156-167. (a)

Yeaton, W. H., & Sechrest, L. Meaningful measures of effect. *Journal of Consulting and Clinical Psychology*, 1981, *49*, 766-767. (b)