Adv. Behav. Res. Ther. Vol. 4, pp. 75–96.
Printed in Great Britain.

0146-6402/82/0275-22$11.00/0

A CRITIQUE OF THE EFFECTIVENESS OF APPLIED BEHAVIOR ANALYSIS RESEARCH

William H. Yeaton
The Center for Research on the Utilization of Scientific Knowledge,
The Institute for Social Research, The University of Michigan, U.S.A.

Abstract — Since its inception, applied behavior analysis has required that solutions to socially significant problems be effective, though criteria for this dimension have remained largely implicit. This paper reviews three explicit technique for determining the effectiveness of behavioral research: graphical, social validational, and cost analyses. The concept of effect size is introduced as an additional means of comparing the effectiveness of various treatment alternatives. Survey data are utilized to support a bothersome implication of this review, namely that the contingency to produce large effects placed on behavioral researchers may actually *decrease* the likelihood that a useful technology of application will be produced. Finally, strategies are offered for preserving the effectiveness of behavioral procedures when existing technologies are disseminated to settings of relevance.

Since the earliest research studies in applied behavior analysis, the primary goal of the field has been to develop a set of solutions to socially significant problems. Research was not merely to be of theoretical importance, it must also demonstrated the effectiveness of treatments. "If the application of behavioral techniques does not produce large enough effects for practical value, then application has failed... Its practical importance, specifically its power in altering behavior enough to be socially important, is the essential criterion (Baer, Wolf and Risley, 1968, p. 96)." Despite the stated prominence given the effectiveness criterion, however, no explicit mechanism was provided for determining whether it had been achieved.

A primary intent of this review is to critically examine three mechanisms (graphic displays, social validation, and cost analysis) which, only through established practice, have attempted to answer the question of whether an effective solution to a socially significant problem has been demonstrated. Additionally, a detailed elaboration of the concept of effect size in applied behavior analysis is presented, with particular attention being paid to those variables which determine the size of the effect produced as well as those factors which influence our ability to detect a wide range of effect sizes. A bothersome implication of the paper suggests that the contingency of producing large effects of clear social importance placed on applied behavior analysts may actually impede progress toward the development of a practical technology of applied behavior change. Finally, two models are presented that emphasize the importance of maintaining large effects as behavior analysts progress from the demonstration to diffusion level of analysis and several possibilities and precedents for the empirical study of diffusion are also introduced.

INTRODUCTION

The first of the mechanisms utilized to indicate that an effective behavioral solution has been demonstrated is the graphic display, a feature present in nearly every applied behavioral study. "Consequently, application, to be analytic, demonstrates control when it can, and thereby presents its audience with a problem of judgment. The problem, of course, is whether the experimenter has shown enough control, and often enough, for believability." (Baer, Wolf and Risley, 1968, p. 94). This paper asserts that the judgmental problem of assessing effectiveness in applied behavioral research cannot be separated from attributions made upon inspection of the graphs appearing in each study. Consequently, much could be

learned about judgments of effectiveness by studying the problem of judgments made from visual inspection.

A second approach upon which judgments are made regarding the effectiveness of procedures utilized in applied behavior analysis is called social validation (e.g. Fawcett and Miller, 1975). Often, judgments regarding the effectiveness of procedures are made by relevant members of society. The formal process is a relatively new one in the field of applied behavior analysis, and the impressions it had made have varied considerably. Social validity has been applauded as an extremely important advancement of the field by Wolf (1978), while Kazdin (1977) has pointed out several of the methodological problems with the process, and Deitz (1978) fears its potential influence may be to suppress the scientific curiousity of researchers in the area. This paper will critique its methodological integrity and offer suggestions for improved methods of validation.

The third approach to the question of effectiveness in applied behavior analysis research is the eminently practical consideration of the costs and benefits of the procedures. Cost-effectiveness and cost-benefit considerations are quite new dimensions upon which to judge the utility of applied behavioral research. The approach is a familiar one in economics and has gained recent stature via the system of planning, programming, and budgeting in President Lyndon Johnson's administration (Schultze, 1968). Application of the principles to psychological research has been relatively rare though the relevance of costing analyses to the field of evaluation has been articulated (e.g. Levin, 1975; Rothenberg, 1975). A critique of costing analysis in applied behavior analysis will allow the reader to assess their promise as an important criterion for judgment about the effectiveness of the procedures presented.

VISUAL ANALYSIS OF GRAPHIC DATA

Visual rather than graphic analysis has long been the choice of applied behavior analysts (Kazdin, 1975b). Michael (1974) has articulated several of the reasons for this choice, and his discussion is a useful departure point in examining judgments of effectiveness made from the graphic displays of data. A central tenet of Michael's discussion is that judgments of effectiveness are best made "out front". Though some degree of "abbreviation" is likely to be necessary in reacting to the raw data of an experiment, graphic displays and descriptive statistics are preferable to computer analyses and inferential statistics in preserving those features of the data from which judgments are made. Sechrest (1976) also maintains that it is a mixed blessing to utilize data analytic techniques that disguise those features of the data from which conclusions are made. He uses the term "opaque" to describe the continuum along which data analytic tests can vary. At the transparent end of the continuum might be the *t* test which is relatively transparent, since those properties of the data which are involved in the statistical test are obvious. At the other end of the continuum, a canonical correlation is extremely opaque since it is not at all obvious which features of the data are crucial to the results of the test. Visual analysis appears to be still more transparent than simple, statistical tests and should allow study of the important dimensions of judgments of effectiveness made by the reader.

A problem inherent to the visual inspection process is the possibility of disagreement between reviewers regarding the extent of effectiveness demonstrate by the intervention. This disagreement is not altogether unreasonable when one considers that judgments of effectiveness are made in concert with the knowledge of the other multitudinous aspects of the research. However, if one is to make any firm conclusions regarding the effectiveness of results based on their graphic protrayal, it is clear that other aspects of the study should not be available to the reviewer when judgments are made.

It is reasonable to take the stance that disagreement among reviewers is a healthy situation and assume that a manuscript should pass the stern test of consensual validation to be accepted for publication by a particular journal. The degree of effectiveness shown must be above the criterion set by the most demanding critics. However, if two, independent reviewers, given only a graphic display, frequently disagree at the mere *existence* of an

experimental effect, the basis upon which judgments of effectiveness are made would appear to be whimsical at worst and subjective at best.

Research on visual analysis

Though the evidence is scanty and only suggestive, it appears that inter-rater agreement between judges can indeed be low. DeProspero, 1976; DeProspero and Cohen, 1979) constructed a set of 36 graphs, systematically varying important characteristics of these graphs and found an average correlation of 0.45 among judges asked to accept or reject a manuscript in which each graph "represents the major results of a manuscript which is being considered for publication". These same judges, chosen from the *Journal of Applied Behavior Analysis* and the *Journal of the Experimental Analysis of Behavior* Board of Editors and/or guest reviewers, rated the degree of experimental control shown on a scale ranging from 0 to 100. The average correlation for these ratings was 0.61. It is critical to note that DeProspero assumed that judgments made in his questionnaire about the extent of experimental control shown and the magnitude of the effect inferred from visual inspection of graphic displays go hand-in-hand. This association was also implicit in Michael's previously mentioned critique of statistical analyses. More explicitly, graphic displays are a primary source for judgments concerning the degree of experimental control shown; graphs showing a "clean" functional relationship between conditions of the experiment are likely to be judged as showing experimental control while graphs showing an ambiguous relationship between conditions of the experiment are likely to be judged as lacking experimental control. And since a judgment concerning the demonstration of experimental control is critical to whether an analysis has been achieved ("An experimenter has achieved an analysis of a behavior when he can exercise control over it."; Baer, Wolf, and Risley, 1968, p. 94), graphic inspection influences *both* the effectiveness and analytic dimensions of applied behavior analysis, and its importance should not be underestimated.

Perhaps even more serious than the lack of agreement between reviewers in their judgment concerning the effectiveness shown in visual displays of graphic data is the lack of agreement between visual and statistical analyses as to whether a significant change in the dependent variable has been produced. Given the traditional claims made by applied behavior analysts that their research findings go *beyond* mere statistical significance (e.g., "If a problem has been solved, you can see that; if you have to test for statistical significance, you do not have a solution." (Baer, 1977a, p. 171)), the lack of agreement is even more distressing.

Jones, Weinrott, and Vaught (1978) have presented evidence bearing on this issue. These researchers chose a non-random sample of 24 graphs from *JABA* and asked 11 judges "familiar with operant experiments ... to decide whether or not a meaningful change in level was demonstrated from one phase to another in each of the graphs." The mean agreement between visual inferences and time series inferences was 0.60. However, their data suggested that as the autocorrelation in the data increased, agreement decreased. It was also the case that agreement between visual and time series results decreased as significance levels increased. Taken together, these two findings are especially important because judges were least reliable (correlation = 0.48) with graphs having high serial dependency and significance levels, and these are the very graphs most likely to appear in the literature.

These findings should be taken as merely suggestive due to several shortcomings in the research. First, the authors state that the effects portrayed in the 24 graphs "had to be sufficiently nonobvious to warrant critical analysis." The vagueness of this sample selection criterion makes it impossible to estimate the generality of the findings. Second, we note that "graphs were simply reproduced directly from the pages of *JABA*". Knowledge of the dependent variable printed on the ordinate may have affected inferences made concerning meaningful change in level. Third, the test items used for visual and statistical analyses were not independent. For example, the three test items from an $A_1B_1A_2B_2$ design, A_1B_1, B_1A_2, and A_2B_2, contain common elements and this renders the statistical analyses non-independent; visual judgments about each of these pairs of conditions were made with all other pairs of conditions in view. Fourth, the numbers for the time series analyses were

apparently obtained by estimating values from the graphs. We cannot be certain about the accuracy of these estimates. Fifth, there may be a confounding of the power of the statistical tests and the extent of agreement between visual and statistical analyses. It may be the case that agreement was least in those instances in which few data points were available for time series analysis. Low power rather than low statistical significance may be the culprit in producing low agreement scores. Sixth, it may be the presence of trend and not the presence of autocorrelation in the data that is associated with the lack of agreement between statistical and visual analyses. Since trends in the data are associated with the presence of autocorrelation (and the authors state that they have chosen experiments "where serial dependency might be evidenced by possible non-zero trend, apparent from visual inspection of the graphs"), the validity of their conclusion about the influence of autocorrelation on agreement scores is open to serious question. The fifth and sixth shortcomings are the kinds of critical flaws so often symptomatic of correlational-type studies.

Factors influencing judgments of effectiveness made from graphic data

There are a host of plausible characteristics of graphs which may influence judgments concerning the effectiveness shown in graphic displays. For example, mean difference between conditions, local and overall trends, and variability within conditions are traditional statistical properties which are likely to influence both visual and statistical analyses. However, when taken together, these properties determine the frequency and extent of overlap between conditions of the experiment and thus may act in concert rather than singly in a visual inspection of the data.

Other possible factors utilized in judgments of effectiveness do not have strict counterparts in the statistical realm. For example, the immediacy of the experimental effect may be an important factor in judgments regarding effectiveness (e.g. Bailey, in press; Ross, Campbell and Glass, 1967). Immediacy of change upon alteration of the conditions of an experiment is a local aspect of what may be construed as a general pattern in the graph, a property that has not been ignored by influential researchers. Some authors have subjectively classified the patterns as "strong" or "weak' indicators of intervention effectiveness (Glass, Willson and Gottman, 1975, Fig. 17), described the legitimacy of inferring effectiveness as "strongest" in one pattern and "totally unjustified" in others (Campbell and Stanley, 1966, Fig. 3), and have listed some varieties of intervention effect patterns (Glass *et al.*, 1975, Fig. 5). Others have attempted to systematically categorize likely combinations of pre-post treatment changes as well as to make qualifying statements about the likelihood of treatment effects (Kazdin, 1976, Fig. 8-3; Jones, Vaught and Weinrott, 1977, Fig. 1). Studies that investigate combinations more complicated than pre-post patterns may be a bit premature in our attempts to gain understanding of factors affecting judgments of effectiveness.

The study of judgments of effectiveness made from visual inspection of graphic displays is a recent endeavor. (See Huff, 1954, for an early, informal critique of graphical displays.) Methodological refinement will be necessary before conclusive statements can be made about the factors influencing visual analyses. Since visual analysis acts as a critical filter through which judgments regarding the degree of effectiveness are funnelled, calls for standardization and presentation of guidelines such as those made by Parsonson and Baer (1978) take on an added significance.

SOCIAL VALIDITY

Social validation refers to a set of formalized methods designed to substantiate the demonstrated effectiveness of an applied behavioral technology. The term is a recent one (Fawcett and Miller, 1975; Quilitch, 1975), though portions of the methodology were practiced prior to its formal inception to the behavioral armamentarium (e.g. McMichael

and Corey, 1969). Historically, its roots can be traced to the position paper of Baer, Wolf, and Risley (1968), p. 96):

> In evaluating whether a given application has produced enough of a behavioral change to deserve the label [effective], a pertinent question can be, how much did the behavior need to be changed? Obviously, that is not a scientific question, but a practical one. Its answer is likely to be supplied by people who must deal with the behavior.

Kazdin (1977) has categorized social validation approaches into those utilizing ratings made by important individuals in the target person's natural environment and those involving relevant norms to which the target person's performance can be compared. An excellent example of the use of both of these social validation techniques is provided in research by Wolf and his colleagues at Achievement Place (Minkin, Braukmann, Minkin, Timbers, Fixsen, Phillips and Wolf, 1976). Components of conversational skills were validated *before* systematic training of these components by having relevant judges rate conversations of nondelinquent junior high school and college students and correlating conversational ratings with the occurrence of behavioral skills used in these conversations. In this way, aspects critical to judgments of superior conversational skills were substantiated as important behaviors to train, and relevant norms were obtained to which ratings of target persons could be compared. Within-subject comparison of ratings before and after training of conversational skills as well as between-subject comparison of ratings anchored in the two normative groups served as persuasive evidence of the effectiveness of training. Substantial change in ratings before *vs.* after training demonstrated that judges could discriminate change between instances of conversation pre- and post-training; normative standards answered the question of whether enough change for significance had been achieved. In fact, the Achievement Place girls' actual performances resembled those of junior high but not college students, demonstrating the importance of the choice of standard in decisions regarding the effectiveness of procedures.

Social validation procedures

Wolf's (1978) conceptualization of social validation procedures is similar to Kazdin's, but Wolf suggests that there are at least three levels: (1) the social significance of *goals,* (2) the social appropriateness of *procedures,* and (3) the social importance of *effects.* In making judgments as to the effectiveness of a given research study, consideration of all three of these levels is relevant. However, this paper will emphasize judgment of and normative standards for the importance of the effects of the research, and only elaborate briefly upon the first and second levels of validation used in Wolf's classification scheme.

Social significance of goals. The validation procedures utilized in the choice of the goals of research focus on careful identification of those behavioral facets of a problem whose modification will lead to a solution. The strength of social validation procedures which correctly identify the relevant behavioral dimensions of a problem in advance of training lies in its virtual guarantee that an effective modification will produce significantly increased ratings by judges. This strategy should lead to a more efficient science. Journals would, theoretically, not be cluttered with articles showing functional control over dependent variables but failing to effect meaningful change in the target population.

In contrast to this ideal state, it appears that current validation procedures rely heavily on a readers' subjective impression (face validity) that the problem and its specification are significant for study. To illustrate, claims for the conduct of school-based research may use personal testimony by a teacher that particular students are not performing at levels appropriate for their age. Institutionalized clients are said to exhibit behavioral repertoires detrimental to their release. Or, an absolute standard may be appealed to, for example, when the stand is taken that since our society values honesty, children who steal are a problem. In each case, we are to be convinced of the social significance of the problem by the authors' argument that change is desirable, or that other researchers have studied the same problem, or that no other researchers have, but should have.

A more empirically based strategy to justify studying a problem relies on the inspection of initial baseline levels in which there is an apparent deficit or excess in performance. To illustrate the inherent weakness of this approach, one can point to published research in which baseline levels of school attendance were well above 90%. If these levels of "deficiency" were presented as *the results* of interventions designed to solve problems of school attendance in other studies, one might be tempted to applaud the degree of success achieved by the authors. More careful scrutiny of target behavior on videotapes or as it occurs naturally might reveal troublesome topographies or behavioral patterns that would argue more convincingly for modification. However, the point remains that there appears to be considerable leeway allowed in establishing the goals chosen for study.

Social appropriateness of procedures. The social appropriateness of the procedures used in applied behavior analytic research is another important facet of social validation methodology. Presumably, a potential consumer's satisfaction with behavioral procedures is critical in determining the likelihood of utilizing a given set of procedures (e.g. Risley, 1975). But what is it about a procedure that influences judgments of acceptability?

One possibility is that procedures align themselves along a continuum of judged severity. In choosing one procedure over another, a consumer may make a subjective ordering of all potential procedures likely to remediate a given problem. The least severe treatment is then the procedure of choice. Kazdin (1980) found the order of acceptability (from most to least) of four treatments for deviant child behavior to be: reinforcement of incompatible behavior, time out, drug therapy, and electric shock, an order that may reflect judgments of severity of treatment. Similarly, Foxx and Azrin (1972) found restitution procedures to be more acceptable to staff than shock or timeout. Whether severity might be judged according to the immediate behavioral effects on the target (perhaps tantruming immediately following initiation of timeout or exaggerated motor and verbal behavior immediately following shock) or according to existing societal values about the use of punishment is conjectural. Finally, the fact that a redirection procedure was judged to be less acceptable than "contingent isolation" by three of five staff in a day-care setting (Porterfield, Herbert-Jackson and Risley, 1976) may be due to the judged severity of redirection as compared to "contingent isolation".

The attractiveness of judged severity as the critical dimension in decisions about acceptability is the implication that the net results would be a set of minimally severe but effective treatments. However, consumer judgments of acceptability are likely to be multiply determined rather than a function of a single factor. For example, would all positive stances for modification (e.g. providing huge amounts of money contingent upon appropriate behavior) be judged more acceptable than, say, minimal amounts of punishment (e.g. a mild tongue-lashing)? Other criteria may be necessary when comparing the full range of procedures from the sets of reinforcing and punishing techniques.

Research from the social psychological literature (e.g. Bickman and Zarantonello, 1978) suggests that it is from the results of research rather than from the procedures utilized that the general public judges the acceptability of the work. When subjects made ratings of several aspects of Milgram's (1963) classic study of obedience, only in those cases in which participants were told that the procedures produced obedience were the ratings substantially less favorable. On the other hand, the effectiveness of four alternative procedures (reinforcement of incompatible behavior, time out, positive practice, and medication) for deviant child behavior did not modify ratings of treatment acceptability (Kazdin, 1981). Clearly, the dimensions upon which judgments of acceptability are made cannot be stated with assurance at the present time.

Another important question to ask regarding the social acceptability of behavioral procedures is: "To whom are the procedures to be acceptable?". Certainly judgments of acceptability may vary as a function of the audience to whom the question has been addressed. For example, procedures may soon be developed which modify the percentage of time that young children wear seat belts when they are passengers in automobiles. From the parents' point of view, these procedures may be totally acceptable, and they will continue to utilize these methods in the interests of the safety of their children. However, from the children's point of view, procedures may be totally unacceptable; the training is boring, there

is no apparent benefit, and their range of activity has been drastically restricted. Which consumer's set of responses is more important? The ideal set of circumstances occurs when a technology is developed satisfactory to all relevant parties. Certainly, if the problems that children create as a by-product of the procedures used also decrease the chances that seat belt wearing will be maintained, the conflict is critical to the effectiveness of the research. Otherwise, current reinforcement contingencies in the culture may dictate whose rights (in this example, parent's or children's) are to be more highly valued.

Social importance of effects. Formal procedures have been developed to assess the social importance of effects produced in applied behavioral research. Typically, bipolar rating scales are given to persons deemed relevant by the research team. These judges are asked to rate the appropriateness of the target behaviors exhibited by subjects along a continuum varying from poor to excellent. For example, Fawcett and Miller (1975) asked judges to rate public speaking behaviors by answering the question "How well would you rate the speaker's overall performance?" on a 7-point scale where seven represents "very good" and one represents "very bad." Such extreme descriptors as these may compress the range of responses made by judges, and raters may tend to respond in the middle of the scale. In Fawcett and Miller's case, it may be unrealistic to ask low-income para-professional staff members to perform up to the standards implied by the anchor "very good." To demonstrate functional control over these ratings will be particularly difficult given that these staff are also unlikely to perform at "very poor" levels initially. On the other hand, using anchors such as "below normal" and "above normal", or "below average for relevant peers" and "above average for relevant peers", may be more realistic goals for the experimenter. Such anchors would also tend to have the effect of spreading out the mean values of responses made by the judges so that more effective change could be demonstrated. At the very least, it appears that the choice of anchors and perhaps the number of scale points may influence the rating of judges.

To illustrate this notion of the inelasticity of rating scales relative to the degree of behavior change produced, four studies were chosen that utilized rating scale techniques. Baseline measures were averaged across all subjects and behaviors. In a similar manner, all behavioral measures taken subsequent to intervention were averaged in the same four studies. Analogously, average pre- and post-*ratings* of behavior change were calculated for these studies. In each instance, percentage of appropriate behavior was used as the dependent variable (Fig. 1). In one case, the 5-point rating scale results were linearly transformed to a 7-point scale. Despite percentage of behavior change as large as 85%, changes in ratings were minimal. The largest change was 2.6 scale points on a 7-point scale. There was no consistent relationship between size of behavioral measure change and size of rating scale change. When translated into these terms, large graphical effects immediately visible to the reader have no apparent social validity.

One study (Frederiksen, Jenkins, Foy and Eisler, 1976) appears to be quite anomalous with respect to the inelasticity shown in the four studies mentioned above. In fact, the congruence of percent appropriate behaviors shown and ratings made by judges was so remarkable that other explanations for the congruence seem likely. In this research, the *same* staff who made ratings of 1, 2, 3, 4, or 5 were also asked to score the presence or absence of four target behaviors and therefore produce percentages of 0%, 25%, 50%, 75%, or 100%. Consequently, the congruence of behavioral scores and ratings appears to be an artifact of the particular procedures utilized. However, there is an interesting possibility which this congruence forces the reader to consider. It may be possible to modify the ratings of relevant staff associated with a behavior change program by having them also score the behaviors that are assumed to change as a result of the intervention. The obvious argument against this procedure is that the rater has simply been co-opted into using the standard of the researcher in judging whether behavior has been changed. One would hope that the research team has chosen exactly those aspects of the behavioral repertoire which are critical to solving the problem, and the rater is, in a very real sense, a substantiation that the correct aspects of the repertoire have indeed been chosen. This discussion suggests that one could hold back particular raters who would be used solely for validation purposes but use other raters to monitor specific behavior change *and* to rate the qualitative dimensions of change. Should

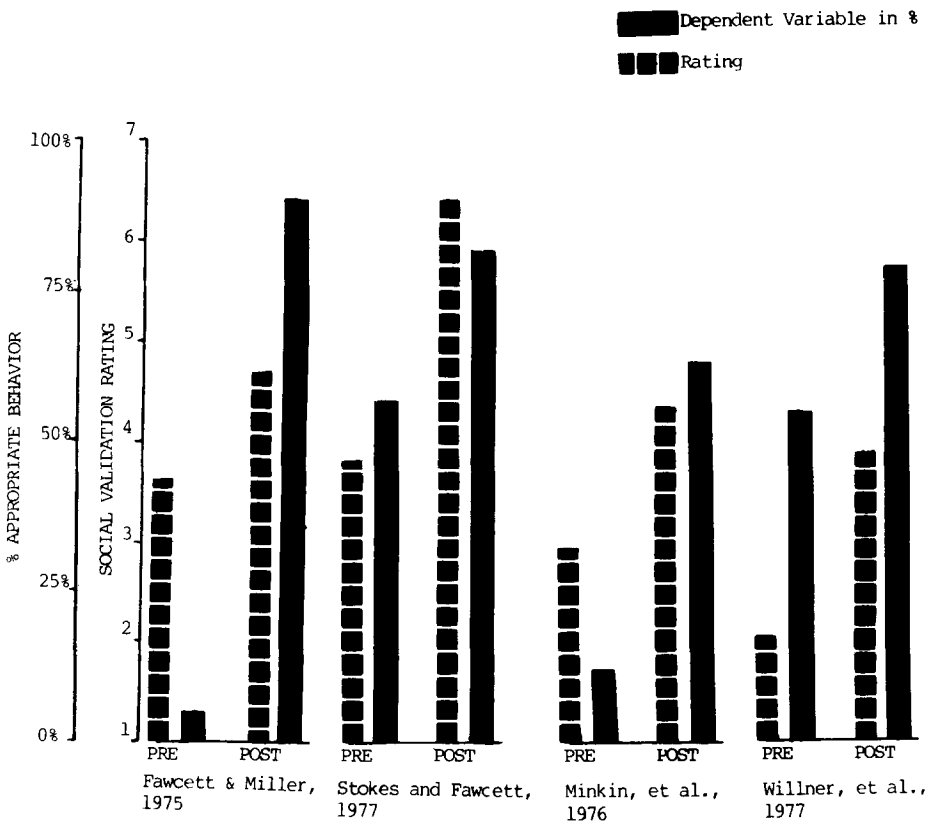

FIG. 1. Average social validation ratings and average dependent variable percentages during baseline and post baseline phases in four *JABA* articles. Three of the social validations utilized 7-point scales; the values of one scale were linearly transformed to corresponding values on a 7-point scale. Values are taken for the next of a study or estimated from graphical displays. Averages were calculated from scores of individual subjects and behaviors or from the particular unit of analysis presented. In cases where decreases in ratings or percentages represented beneficial changes, these values were replaced by their symmetric counterparts (e.g. a value of 40% was replaced by 60% and a rating of 2.3 on a 7-point scale was replaced by a rating of 5.7). In Minkin *et al.* (1976) raw scores were converted to percentages. In Willner *et al*, (1977), percentages for positive teaching-parent behaviors were utilized in the calculations.

these latter judges who have made higher ratings than expected also be more likely to change their pattern of interaction with the target in ways which may serve to maintain the behavior change produced in the research, an apparent artifact may prove to be a viable option for programming generalization.

Critique of social validation procedures

There are other possible procedures which may be used to facilitate the discriminations of change made by relevant staff prior to and subsequent to training. These procedures do not socially validate the effectiveness of the results produced. Rather, they take the form of social validation methodology which actually changes the effectiveness of the results. One possibility is to ask the rater what a person competent or non-competent in a particular area might be doing prior to their completion of the rating scale. This questioning may serve to sensitize them to aspects of behavior which they perceive to be important to skillful and nonskillful performance without cueing them as to those aspects which the researchers have chosen to train.

Another possibility, one which clearly would co-opt the rater as an independent validation of the effectiveness of the procedures but which would have the advantage of greatly increasing the chances that the rater would discriminate important differences, would require verbal descriptions and/or videotaped examples to illustrate the anchors of the

rating scale. Having the relevant behavior described to them or actually viewing instances of the extremes of behavioral performance should tend to increase the judges' ability to discriminate differences in performance. In principle, a series of brief tapes that show less and less or more and more of the behavioral complex of interest could be viewed, much in the same manner as the method of limits in psychophysical experiments (cf. Nachmias and Steinman, 1965). At that point of reliable discrimination, one would be able to measure the degree to which behaviors chosen for training need to be present for a given judge to rate differences in responding as important. By repeating this procedure with several raters, a range of occurrence values would be produced which would yield a minimum value at which discrimination is likely to occur for even the most insensitive rater. These procedures would not eradicate the problem of variability across judges (Bailey, 1978) but would at least determine a minimum standard of effectiveness of training. Phillips, Phillips, Wolf, and Fixsen's (1973) discovery that a 75% criterion of appropriate behavior was associated with high ratings, while a criterion of 50% produced substantially lower ratings illustrates the critical importance of parameters chosen. And though these procedures would change the value systems of the raters, the likelihood that these raters would behave differently in the presence of target subjects would be increased.

One difficulty in utilizing rating scales for social validation purposes stems from the inadequancies of the procedure as the *sole* means of assessing the effectiveness of treatment results. By itself, it is not particularly insightful to know that an average rating of 5 on a 7-point scale has been achieved. Ultimately, the worth of applied behavioral procedures will be compared to procedures developed by other orientations (e.g. humanistic). If an alternate orientation has consistently produced average ratings of 6 on a 7-point scale, one's enthusiasm for the procedures associated with an average rating of 5 will be quite different than if the alternative orientation has only been able to produce average ratings of 4.

The point is that it may be meaningless to attempt to make *absolute* judgments about the effectiveness of research, and, given that social validation techniques are relatively novel, there is only a small data base upon which to make judgments between studies in the same area but across orientations, or across areas using the same orientation. Comparability of scales (e.g. same anchors used, same number of points on the scale) is one obvious prerequisite to make these sorts of comparisons. Though the difficulties are complex (cf. Sechrest and Yeaton, 1981b), such procedures may be necessary for meaningful comparisons between treatments (O'Leary, 1977).

Behavioral researchers may also wish to invest their collective energies in studying other facets of the behavior besides frequency of occurrence since these facets may be critical to the judgmental process. Perhaps there are topographical characteristics of the behavior (e.g. response amplitude) or predictable patterns (e.g. immediacy of effect) which act singly or in concert to make it more likely that a judge will discriminate a difference in pre-post performances. As Wolf (1978) has mentioned, we cannot be certain that high ratings are a sufficient indication that the critical dimensions of the behavior have been identified. In traditional terms, construct validity may be lacking (cf. Cook and Campbell, 1976).

In his critique of social validation procedures, Kazdin (1977) calls attention to the possibility of utilizing traditional psychometric criteria to evaluate the rating scales approach to social validation. For example, the sole reliance upon face validity for selection of the dimensions to be rated is problematic, while other criteria (e.g. convergent validity) offer promise for assessing the social significance of both the goals and the effects of behavioral techniques. What is being suggested here is that multiple specifications of the behavior being studied be utilized *within a given study;* in other words, at least two different definitions could be developed for each behavior measured. This principle of convergent validity (Campbell and Fiske, 1959) is not new but has typically been utilized for the sake of convenience. To illustrate, Lippencott (1978 used a traditional scrape and weigh procedure to validate a much more convenient rating system based on rater judgments of how much food in different categories was left on the plates of school children. Jacobs (1979) used the unwieldly method of weighing newspapers to validate a more convenient technique of simply measuring the height of the newspapers being recycled.

The logic behind using two different definitions in a given research study lies in the

demonstration that the effects produced are not simply a function of the, perhaps arbitary, choice of definition and consequent measurement format. In some instances, it may be particularly convincing to develop two *independent* sets of behavioral definitions. For example, if one definition of "warmth of a therapist" was developed by a group of clinical psychologists, the believability of the effectiveness shown using measurements based on the behavioral definitions would be greatly enhanced if the other measurement system's findings were obviously consistent. When procedures are effective regardless of the definition chosen, considerable credibility is given to the results, especially when any bias in the development of the definitions by clinical psychologists would be in the direction *opposite* that of the behavioral psychologists.

Norms as social validation standards

Normative approaches to assist in judgments of the effectiveness of applied behavioral research have been used relatively frequently. Chapman and Risley (1974) relied upon the number of pieces of litter observed on selected middle-class lawns as a standard against which to judge the effectiveness of their litter-reduction procedures. Walker and Hops (1976) used normative peer data as a means of assessing the worth of procedure designed to increase the percentage of appropriate behaviors in the classroom. As Walker and Hops mention. such standards allow important qualifications to be made about the generality of procedures across time. Decreasing trends during follow-up which parallel decreasing trends in normative subjects are less indicting of the quality of procedures. Studies such as White's (1975) of the natural rates of teacher approval and disapproval in grades 1 through 12 provide appropriate methods to which programs designed to change rates of approval and disapproval might be compared. The previously mentioned work by Wolf and his colleagues (Minkin et al., 1976) also illustrates the use of normative data as a supplement to standard research methodology. Ciarlo (1977) has used a range of community scores on such dimensions as psychological distress, non-productivity, drug use and consequences, and client satisfaction as standards to measure the effectiveness of program outcomes. It is interesting to note that the upper point of the acceptable range of Ciarlo's band is at the mean score for the community. Clearly, his goal is not to bring clients to a super-state of functioning, and this choice of goal is critical in determining the success of the program.

The preceding examples of normative standards are all similar in their extraction of data upon which the norms are based; all use a logically relevant, (same sex, similar age, similar SES) locally derived group to compare experimental effects to. Across-study norms are also promising standards to use as yardsticks to measure the worth of research. For example, Glass and his colleagues (Smith and Glass, 1977; Smith, Glass and Miller, 1980) have aggregated findings from several hundred therapy studies in an effort to determine the degree to which treatment groups are superior to control groups. The results of any given therapy study could be evaluated according to the point in the distribution where it would fall (e.g. at the 37th or 97th percentile).

The logic of this approach is not new as effects of several studies are often presented in tabular or graphical form (e.g. O'Brien, Hamm, Ray, Pierce, Luborsky and Mintz, 1972). However, this technique is a promising means of assessing the worth of interventions and offers an improvement in the traditional practice used in review papers of simply counting those which do and do not yield statistically significant results (e.g. Kilmann and Sotile, 1976). Rather than basing conclusions on the majority "vote", these normative methods weight the effectiveness measure of each study and estimate important parameters such as the mean and standard deviation from the sample of effect sizes.

Normative standards, however, may be inadequate in several ways (Sechrest and Yeaton, 1981a). As Van Houten (1978) has commented, in classroom studies collateral data on *both* teacher and student behavior may be necessary in order to place value on a particular teaching practice. Different standards may be needed for children from diverse backgrounds with differing problems. Distinct situations may call for varying amounts of, say, teacher attention. Furthermore, the consideration in choosing an appropriate norm may be more

difficult than at first glance (Van Houten, 1979). What, for instance, would be the appropriate norm group for "shy" preschoolers who have been taught to increase the frequency of their social interactions? What if the normative standard is itself deficient, as in the incidence of passengers in automobiles who consistently wear their seatbelts? Would procedures have been effective if they had doubled the incidence of consistent seatbelt users in a community? Unfortunately, these difficulties suggest that the formulation of norms will be quite time-consuming since each locality as well as specific subject population and situation may well demand its own normative standard.

COST ANALYSIS

Though cost analyses are a new feature in applied behavioral research, their influence has been longstanding in such diverse areas as emergency medical services (e.g. Acton, 1977), traffic safety (e.g. Lave and Weber, 1970; Little, 1968; Valavanis, 1958), public health (e.g. Schramm, 1977), and political decision making (e.g. Schultze, 1968). Comprehensive surveys (e.g. Prest and Turvey, 1965) and books of readings on pertinent issues (e.g. Niskanen, Harberger, Haveman, Turvey and Zeckhauser, 1972) are only a portion of the voluminous literature on the subject. Recently, articles by Levin (1975) and Rothenberg (1975) have elaborated upon cost-effectiveness and cost-benefit analyses from a psychological perspective.

A rationale for costing procedures in applied behavior analysis emphasizes the consumer acceptability of procedures. At a demonstration level of analysis, dollar value assigned to costs and benefits are less important than showing that appropriate modifications are possible. However, any move in the direction of broad dissemination of behavioral methodology must take the financial attractiveness of the procedures into account. Finally, costs and benefits are very practical and very powerful facets of comparisons between behavioral and alternative treatments.

A survey of costing techniques

To illustrate the increasing important of costing procedures in applied behavioral research, a systematic survey of the first 10 volumes of the *Journal of Applied Behavior Analysis* was conducted to determine the extent and expressed purpose of cost considerations in full research articles. Reliability was calculated on an article-by-article basis; the occurrence reliability was 72%, while the non-occurrence reliability was 98%. Dollar estimates were found in only 25 of the 409 articles (by both raters) during the 10 year period. Sixty percent of the cost citations were cited in the last two volumes; the first eight years only produced 10 citations, evenly spaced across this interval. The most frequent use of costing techniques involved efforts to quantify some aspect of treatment cost and nearly half of these articles also estimated the costs of an alternative program from actual data produced in the study or, more often, from archival records. Results from the survey support the subjective impression of an upward trend in the use of cost figures in applied behavioral research. This trend parallels the recent emergence of community application of behavioral technology (Fawcett, Mathews and Fletcher, 1980; Glenwick and Jason, 1980), an area in which cost considerations may be particularly relevant.

Shortcomings of current costing techniques

Current cost analyses utilized in applied behavior analysis lack the technical sophistication of costing analyses used in economics in which there exists established criteria to determine the merits of costing efforts. None of the 25 analyses in the above survey directly utilized the concept of opportunity cost (the cost of foregoing the most attractive alternative), though other alternatives were occasionally given a dollar value. Seldom were

any monetary estimates made of the many benefits accruing from programming. Even the exemplary evaluation by Paul and Lentz (1977) of hospital, milieu, and social-learning treatments of chronic mental patients utilized a cost-effectiveness rather than a cost-benefit analysis. Cost and benefit streams (dollar value across time) were not reported so discount rates could not be applied. Consequently, investment criteria such as internal rate of return, net present value, and benefit-cost ratios, whose purpose is to make a point estimate of the worth of programming, were also not applicable. These criteria are perhaps prematurely stringent to invoke at this point in the development of costing procedures in applied behavioral research, but it is unsettling to find naive estimates of the worth of research based on samples of a dozen or so being generalized to community-wide efforts without any consideration of the costs incurred at a systems level of implementation.

Finally, as applied behavior analysis expands the domain of its influence to larger numbers of individuals, distributional considerations will be emphasized more frequently. Weisbrod (1968, 1972) has produced evidence indicating that the demographic characteristics of the recipients of benefits may be an important factor in choosing between programs; two programs with equal cost-benefit ratios may not be judged as equals if one program distributes its benefits to a larger portion of, say, low income families than the other. Self-interest must also be considered, as when politicians favor programs more favorable to their own constituents or when staff members opt for alternatives that distribute a larger proportion of benfits to themselves rather than to clients. Irregardless of inherent problems, however, the lure of accountable procedures attractive to the consumer seems to ensure that considerable effort will be expended to develop more sophisticated costing analyses in applied behavior analysis.

THE CONCEPT OF EFFECT SIZE IN APPLIED BEHAVIOR ANALYSIS RESEARCH

Whichever combination of techniques is chosen to demonstrate effectiveness (e.g. graphic displays, social validity, or cost analysis), the clear emphasis of the field of applied behavior analysis is to produce socially important effects of sufficient size to be discriminated by both lay and research audiences. Ironically, effect size, per se, is not a concept frequently discussed by applied behavior analysts.

In psychological research, effect size refers to the quantification of the difference between the means of independent conditions of groups in the experiment (Sechrest and Yeaton, in press). Usually this mean difference is standardized by dividing by an appropriate standard deviation (e.g. the control group or the pooled experimental and control group standard deviation). For purposes of this presentation, the difference between conditions, as in the difference between the means of baseline and treatment conditions in the within-subject designs of applied behavioral research, will be utilized as the definition of effect size. Though this difference is easily obtained, other dimensions of the data mentioned previously are likely to influence *judgments* concerning the size of the effect calculated from the difference between means. And it is in the spirit of clear functional relationships that applied behavior analysts are likely to make judgments about the size of the effect shown.

What is a big effect?

Applied behavior analysts have not defined what they mean by a big effect. Baer (1977b) has addressed the issue indirectly, insisting that applied behavior analysis is the study of powerful variables, i.e. "turning away from the detection of weak variables." He refers to differences between conditions in other areas of study as "much smaller" and "less consistent" than those produced in applied behavioral research. Avoidance of Type I error is given considerable preference to the avoidance of Type II error (Baer, 1977a) though this stance is simplistic and not universal (e.g. Nagel and Neef, 1977). Since this notion of big

implies a comparative standard, we need a methodology that will allow us to make judgments that one orientation to research produces consistently bigger effects than another.

How might this question of comparing effect sizes in different studies or treatments be answered? Could we compare the data of several studies in each area? Are subject populations comparable? How do we equate the strengths of the treatment given? Are initial levels of performance the same? Are treatments given for approximately equal periods of time? By staff with equal qualifications? Perhaps the answer lies in conducting studies that make this comparison explicitly. Surely an empirical answer would satisfy those members of both orientations. But who will conduct the research? Should "equally competent" members of each orientation be given randomly assigned halves of a subject pool and be asked to optimize their efforts at modifying the problem being presented? How would the cries of bias be allayed if a member of a given orientation utilized both of the rival approaches? Would "straw men" be set up, only to be destroyed? Would both procedures be given with equal care and enthusiasm by members of a given discipline? Clearly, the issues are not simple and statements of superiority will have to be backed with data, but what data and from whom are not at all clear.

One possibility for arguing that a big effect has been achieved might be to use existing, agreed upon standards to validate the changes produced in the dependent variable. Surely, we would judge a safety program that saved even a single life as demonstrative of a big effect. To illustrate further, Gori and Lynch (1978) use the adjective "tolerable" to describe the risk associated with smoking certain brands of cigarettes in moderate rates, implying that these levels of consumption produce little danger to the consumer. Presumably, modification which reduce smoking rates to the tolerable risk category could legitimately be termed big if this line of argument is accepted. A program that changed the blood pressure of patients currently at risk to a level of risk which had been substantiated by mortality tables to be significantly less dangerous in terms of life expectancies might reasonably be termed a big effect.

A common difficulty lies in the disconcerting fact that large and reliable change may not be socially significant, as in the case of large changes in blood pressure that leave subjects in the same at-risk category they were in prior to treatment. Ironically, much smaller effects may be clinically significant, as would be the case when changing the percentage of appropriate street-crossing behavior of a young children from 75% to 85% makes it extremely unlikely that a dangerous confrontation with a motor vehicle would ever occur, while changing their level of appropriate pedestrian behavior from 20% to 60% may be inconsequential in terms of their actual probability of being killed or injured. Effects of procedures which produce modest gain in the rate of appropriate responding but trap participants into more encompassing sets of constructive activities (e.g. Baer and Wolf, 1970) are also of this genre.

There is yet another sense in which small effects, in absolute terms, may be practically significant. In education, for example, no single bit of information is critical to a person's future well-being. Rather, it is the gradual, steady accretion of knowledge which is important. Likewise, interventions which produce monotonically favorable trends are to be salvaged as useful. However, one must assume that effects are monitored over many sessions or the benefits of treatment are not likely to be apparent. It is also problematic that such weak but cumulative effects are vulnerable to other plausible explanations for their functionality.

The relativity of effect size

Applied behavior analysts tend to talk about the production of big effects in an absolute fashion when they would be spoken of more appropriately in relative terms. Surely, one could not expect to alter the car-pooling behavior of adults to the same extent as the language behavior of young children. It is probably a good deal easier to teach young children to name their colors correctly than to share their crayons with other young children. It may be the case that certain skills, such as learning a foreign language, are more easily learned at an

earlier than at a later age, though the concept of "readiness" is often utilized to remind us that efforts to teach particular behaviors may be inefficient of effort if attempted prematurely. Even with complete contingency control over behavior, it is unlikely that change can be accomplished with equal ease with all behaviors.

There appears to be many qualifying dimensions which could be considered important in categorizing an effect as big. These dimensions are illustrative of the fact that though mean difference is an objective criterion upon which to judge the size of an effect, other criteria may have considerable subjective influence. Effects produced with little cost and at minimal effort are probably judged to be "bigger" than the same effort produced with considerable cost and maximal effort. Effects produced with some kinds of problems (e.g. criminal activity) and subjects (e.g. the elderly) probably suggest a greater effect than the same results with other problems (e.g. tantruming) and subject populations (e.g. kindergarteners). We are likely to call effects at the diffusion stage of research "bigger" than equal effects at the demonstration stage, perhaps because aspects of the experimental situation (e.g. wider ranges in abilities of the target subjects and change agents) and knowledge of the area make it less likely that the same degree of enthusiasm and contingency control can be mustered at the diffusion stage. Effects associated with larger numbers of subjects may be deemed "bigger" than those associated with considerably smaller sample sizes, perhaps because we judge the degree of attention to individuals that is possible to be proportionately less with larger number of subjects. It would also be natural for us to judge effects as "bigger" when a single independent variable is used rather than when several independent variables are utilized simultaneously (a treatment "package"). Effects that last for relatively longer periods of time are likely to be seen as "bigger" than those which have limited "holding power". Those effects produced in the field as opposed to the laboratory are likely to be judged as "bigger" due to the smaller degree of control we are likely to have over all influential variables (cf. Sidman, 1960). Effects produced by programs utilizing paraprofessionals are likely to be judged as "bigger" than those using skilled behavior modifiers. Effects resulting from sessions of short duration and over a limited number of time periods are probably going to be seen as "bigger" than effects resulting from sessions of long duration and over multitudinous time periods. A dependent variable whose behavioral conceptualization and resulting reliability of measurement may be quite difficult to attain (e.g. fathering) would likely be associated with effects judged to be "bigger" than those produced by a relatively simpler conceptualization (e.g. verbalizations) where high levels of reliability and construct validity are easier to establish. Effects of interventions that modify multiple target behaviors simultaneously may be called "bigger" than those same interventions that improve smaller numbers of target behaviors.

The point, a rather long-winded one, is that judgments of effect size are not absolute but are instead quite relative to the unique aspects of an experiment. To speak of them otherwise is to ignore these and other unmentioned dimensions of research which contribute to this non-absoluteness.

It is quite feasible to investigate empirically the influence of these qualifying dimensions on the judgment of effect size. For example, if one wished to study judgments of effect size as a function of the treatment given, a between-groups comparison could be made by instructing one group of judges that the graph portrayed the effect of independent variable X on dependent variable Y and the second group of judges that the same graph portrayed the effect of independent variable Q on dependent variable Y. Or, one could show two groups of judges the same graph and tell one group that the effect of variable X on variable Y was being examined for subject population P; the other group might be told that the results shown were true for subject population O. Studies of this sort may themselves yield small effects since the influence of any *one* of these facets on the judgment of effect size might be minimal. However, when several of the dimensions are contained in the same experiment, their interactive effect might well be substantial.

As applied behavior analysis begins to move towards community applications (Yeaton, Greene and Bailey, 1981) and into systems containing larger numbers of individuals (e.g. Barber and Kagey, 1977), it is plausible to expect that procedures will not influence the behavior of all persons to the same extent. Another dimension of effect size may well be the

percentage of people of a certain type who are influenced by a given program. For example, Children's Television Workshop assumed from the beginning of their efforts to develop high quality TV programs for children that these programs should not selectively influence children in differing socio-economic strata (Lesser, 1974). Society's implicit weighting of greater value to effects produced with children having lesser educational opportunity may have considerable influence on judgments of effectiveness.

Effect size and power in applied behavior analysis

The concept of power (e.g. Cohen, 1969; Feldt, 1973) is quite familiar to researchers using designs in which statistical analyses of results are common. The power of an experiment refers to the probability of finding a difference between groups and will systematically vary as a function of sample size, probability of Type I error, and effect size. Increasing sample size and effect size as well as increasing the probability of Type I error all serve to increase the power of an experiment. We might logically ask which experimental decisions would tend to increase the chances of detecting differences in applied behavioral research. Certainly, increasing effect size will increase power in both orientations. Though there is no strict counterpart in applied behavioral research, any decision, either implicit or explicit, that would increase the probability of a Type I error would lead to greater power in both orientations.

Perhaps the most fruitful analogues consider the relation of sample size to the power of an experiment. In behavioral research, concern regarding sample size enters in two ways: first, with respect to the number of data points gathered, and, second, in terms of how many persons have been made available to study. Large numbers of data points during baseline and treatment phases may increase the sensitivity of the analysis to detect intervention effects despite the existence of cycles or trends in the data. The accuracy of reliability estimates both between observers and within the same observer at different periods of time can be greatly enhanced by the availability of substantial numbers of data points. Complex patterns of effects or weak effects, though replicable in, say, the several legs of a multiple-baseline design, are unlikely to be discovered with small data sets. This phenomenon is analogous to the relatively weak power of tests for interactions within cells of ANOVA designs. In the second, more traditional sense, when we speak of sample size as the number of participants in a study, it may be the case in applied behavioral research that we are *less* likely to detect differences when larger N's are chosen.

Early criticism of behavioral research for the study of individual and small group cases may have been based on a correct folk wisdom-intuition that it is more difficult to obtain desired results with larger N's, thus implying that a technology based on limited numbers of subjects would become invalid when larger sample sizes are utilized. Belief may have been based in fact. The mechanism for this "washing out" of treatment effects may be the failure to implement the independent variable as planned with all study participants; complete contingency control may become less possible with increasing sample sizes.

Power loss due to degradation of treatment. The notion of a continuum along which implementation varies may be quite critical in understanding why programs fail when implemented in realistic settings (e.g. Charters and Jones, 1973; Hall and Loucks, 1977). Some researchers (e.g. Freeman, 1977) consider this deficiency to be the primary reason for lack of impact in evaluation studies. Boruch and Gomez (1977) have discussed the possibility of developing "a more informative theory of statistical power for evaluations" and consider the degree to which program is implemented as intended as one of the critical determiners to of the power of the evaluation to detect differences. Intensive treatments can be degraded in any number of ways and one cannot assume without careful monitoring that adherence to treatment plans has been a reality (Sechrest and Redner, 1979; Yeaton and Redner, 1981).

When programs are multi-faceted, it becomes critical to know whether all components are equally demanding of effort to ensure exact implementation. When large training staffs are involved, it becomes very cost ineffective to use an aspect of a training program that is not associated with positive results. Worse than being inefficient, negative side effects may

follow from the implementation of more than a minimally sufficient set of procedures. For example, professional staff may become significantly less careful in the degree of attention given to detail if they have to spend fifteen rather than ten minutes each day over a period of months. If shortcuts are made by staff to reduce programming to a more comfortable ten minutes each day, the research team may find that it is exactly those program elements most critical to success which have been dropped. This is less likely to occur if components are, at least, logically related to outcome, but it may be necessary to warn staff that failure to utilize each of the aspects of the program could render lesser efforts useless. These aspects may include more subjective factors such as the degree of enthusiasm and warmth shown, which turned out to be the case in the replication of the Achievement Place model (Wolf, 1978).

THE PLAUSIBLE EFFECTS OF THE "BIG EFFECTS CONTINGENCY"

Though not stated in an explicit manner, the implication is often made that behavioral effects are larger than those produced by other orientations. For example, consider the following statements: "... the behavioral approach often seeks large behavior changes." (Kazdin, 1975a, p. 23); "However, some of the effects produced after a generation of the experimental designs were strong and robust, singly or in groups." (Baer, 1975, p. 17); "As a result, they (individual-subject-design practitioners) learn about fewer variables, but those variables are typically more powerful, general, dependable, and — very important — sometimes actionable." (Baer, 1977a, pp. 170–71). It is a central thesis of this paper that the contingency to produce large effects placed on researchers in the field of applied behavior analysis can have very real effects on the kinds of research accepted into the field's journals. More specifically, if research demonstrating large effects is paid homage to, it is natural that such demonstrations would appear. While it is entirely possible that such a stricture simply acts as a sieve, selecting out those research endeavors which are inferior in design and conduct and whose treatments are potentially less powerful or implemented in less thorough ways, the absence of data permits other equally plausible possibilities.

One such possibility assumes a scenario in which researchers make a multitude of informal decisions influencing the magnitude of the experimental effects shown. If one believes in contingencies and a research audience sensitive to them, it is implausible to conclude that any research decisions made would intentionally jeopardize the chances of producing results which are publishable in applied behavioral journals (i.e. results demonstrating big effects). Arguing a case for this possibility is exactly that, an argument, though the case is strengthened with the knowledge that a similar phenomenon has occured in other fields of study. Campbell (1975) has advanced the notion of the corruption of those particular measures which are chosen to make a case for or against a particular social intervention. By corruption Campbell means that a given choice of dependent measure would predictably be invalidated as a veridical indicator of change. If workers are paid on a strict piecework basis, one would expect the quantity of their work to increase at the expense of quality. If departments of rehabilitation are rated according to the number of successful rehabilitations (clients placed and studying on the job for at least 60 days), one would predict the indicator to increase. However, the increase could well be due to the choice of less severely disabled clients being placed in work settings less approximate to their maximum level of functioning. The possibility of such corruption in applied behavioral research is, then, by analogy, plausible and sensible.

The manner in which this hypothesized corruption occurs and its potential impact is conjectural. If one accepts the premise entertained in a previous section of this paper that general predictions can be made regarding the size of experimental effects with small vs. large numbers of subjects. expert vs. lay trainers, demonstration vs. diffusion levels of programing, etc., then it is possible to archive information from journals to indicate any potential bias in choice of aspects of an experiment which may indeed influence the degree of experimental control shown.

To assess the possible influence of these "incidental" aspects of applied behavioral research, a systematic survey of Volume 10 of the *Journal of Applied Behavior Analysis* was

conducted to investigate the characteristics of full research articles published. In 30 of the 58 articles (52%) (sample size could be determined, in all but one article), 1–4 subjects were utilized. Children were used in 41 of the 59 studies (69%) in Volume 10. Of 59 studies, 45 (76%) were conducted in labs, schools, and institutions. Thirty-three percent of the percent of the studies (18 of 54 studies where an intervention was made by a person rather than by, say, an apparatus or sign) involved the experimenter, a confederate, or a therapist. Training was given individually (i.e. one-to-one) in 42 of the 59 studies (71%) in which it was possible to determine from the article the staff-to-client training ratio. In the other 29% of the studies, training was given in groups or both individually and in groups. For the above categories, the range of reliability of two independent raters was 72–90%; the overall mean reliability was 82%.

Taken singly, none of these results is particularly indicting. However, it is entirely possible that the magnitude of experimental effect could be appreciably changed when several of those factors appear together in the same study. This is not an argument against the internal validity of behavioral research, rather, it is an argument that the results are most likely to be generalizable to situations in which the same characteristics are present (*external* validity). Thus, we are left to consider the extent to which we have produced an *applied* technology of behavior change. To the degree that the data suggest that we have worked with limited numbers of subjects, typically young and in rather restricted environments, utilizing highly-skilled trainers in one-to-one situations, is exactly the extent to which the generalizability and practicality of our technology is called into question.

TOWARDS A MORE APPLIED TECHNOLOGY OF APPLICATION

To this point in its development, applied behavior analysis has emphasized the design and analysis of technologies of behavior change without asking if these behavior change strategies can be disseminated without destroying their demonstrated effectiveness. An applied technology of application will focus on analyzing those variables which are likely to determine the effectiveness of currently validated procedures implemented in settings of relevance. In this vein, it is difficult to imagine a more critical variable than the degree to which treatment is implemented as planned. We must determine, first, how procedures that have been experimentally validated are implemented in applied settings and then, if the results of the implementation are sufficiently "washed out" for practical value, designing a solution to avoid the "watered-down" effects.

The importance of monitoring the intervention

Previous mention has been made of the importance of monitoring the implementation of the independent variable in treatment programs, primarily so that firm conclusions can be made concerning programs appearing to be ineffective. However, if firm statements are also to be made concerning the magnitude of effectiveness of *successful* programs, it is equally critical to determine the extent to which programing has been delivered as intended in these instances. Otherwise, our penchant may be to make the assumption that a given treatment has been implemented exactly as planned during all sessions and to infer that the magnitude of the treatment effect is the maximum possible. Only under conditions of complete implementation are meaningful comparisons of the effectiveness of different approaches to the same problem attainable (Yeaton and Sechrest, 1981). However, is it realistic to assume that a behavioral approach to a problem such as hyperactivity can be implemented in the exact "dosage" specified and as consistently as a 10 mg tablet of methylphenidate, twice a day (cf. Shafto and Sulzbacher, 1977)? Analogously, it may be entirely plausible that the demonstrated superiority of, say, a behavioral approach as compared to a humanistic approach for a given problem is due to the behavioral approach's relative unlikelihood of being degraded when utilized in an applied setting. More specification of method may be possible with a behavioral approach. Or, perhaps the extent of effort necessary to state

exactly what is meant by particular non-behavioral orientations would make these approaches cost ineffective, especially if the specification has to be practiced rather extensively before it can be used with exactness in an applied setting by non-Ph.D. staff members. A comparison of the effectiveness of behavioral to nonbehavioral approaches to the same problem is premature until the time when more careful specification of the degree to which the independent variable has been implemented as planned becomes routine practice.

Unfortunately, standards in the behavioral literature do not appear to require monitoring of the independent variable. For example, in Volume 10 of the *Journal of Applied Behavior Analysis,* only 16 of the 59 regular research articles (27%) supplied data monitoring the degree of implementation of the independent variable. However, more than half (11 of 16) of these studies did supply reliability estimates on the implementation of the independent variable.

Precedents and possibilities for diffusion models

Appropriate models for the diffusion of innovative technology in applied behavior analysis are an infant area of research concern (Stoltz, 1981). In the early stages of a behavioral technolody of education, an unsatisfactory diffusion model involved the development of short-term workshops to train teacher's correct usage of behavioral principles in the classroom. The dream of simply presenting principles and allowing practice in behavioral techniques to groups of teachers who would then return to their own classrooms equipped with an arsenal for accelerating appropriate and decelerating inappropriate behaviors proved to be illusory. Researchers (Fairweather, Sanders, and Tornatsky, 1974; Stein, 1975) have cautioned their colleagues to beware the potential dangers of consumer misuse of procedures. Ineffectual procedures may be abandoned easily and disparaged promptly regardless of the locus of fault for their misuse.

Initial conceptualizations of more effective diffusion models can be created by careful scrutiny of those procedures which are systematically utilized in our society to dispense available goods and services. For example, if you experience problems with the family automobile, you may consult a friend or service manual for tune-up specifications. More serious problems usually require the attention of a mechanic. Runny nose, sore throat, and minor aches and pains can be self diagnosed and remedied by a trip to the local drugstore or medicine chest. A visit to the family doctor is appropriate for broken bones, serious skin infections, or shortness of breath. Self-help materials are readily available in public libraries or bookstores for routine vocational and marital adjustment problems. A therapist's help may be necessary when stress interferes with day-to-day responsibilities.

The models of diffusion inferred by these examples are of either a self-instructional or expert orientation. In each case, a trained specialist is required for those problems likely to be judged more severe and consequently in need of stronger treatments. The self-instructional treatments, as distinguished from those administered by experts, share the common features of having less potential for irreparable damage if misused, being less demanding of immediate attention, and allowing a relatively higher tolerance of error in implementation.

Thus, magnitude of problems tend to align themselves along a continuum of judged severity. Strength of treatments chosen as solution tactics will then covary in direct proportion to the best estimates of effect size which our collective experiences determine to be necessary. Somewhat arbitrarily (though perhaps with a cost-benefit intuition), society appears to encourage expert intervention with more serious problems and allows self-instructional participation with less serious problems.

Traditionally, the diffusion model preferred by psychology appears to be the expert variety, regardless of the judged severity of the problem. Prototypically, a psychologist is likely to be approached about a particular problem, offer a best guess for solution, and consult with the client in an on-going capacity. This model has not generated important research questions regarding its adequacy or efficiency.

The self-instructional model, however, not only offers a viable alternative to the expert approach but also generates a host of questions that may be answered empirically. In applied

behavior analysis, we must begin to ask whether our research efforts should move in the direction of validating treatment packages that could be used independently by relevant consumers or whether we should maintain our status as expert consultants. Considerable expenditure of research effort may hinge on the answer to this question. If the answer is in the affirmative — we do wish to empirically investigate the diffusion of self-instructional technology — it is worthwhile to speculate upon the course of this tack.

What may be needed are behavioral prescriptions, analogous to prescriptions given in medicine, that would accompany descriptions of procedures. These prescriptions would state important restrictions and regulations that should be followed carefully and the likely results if procedures are not followed closely. For example, a warning may be provided: "These procedures will reduce the incidence of tantruming in your child if given in the amount and manner prescribed. If given in a smaller amount, at best, the procedures will prove ineffectual; at worst, you may have escalated the strength of treatment necessary to modify this behavior in the future. Damaging side effects such as failure to follow directions you give to your child may result from departure from the procedures described." (Yeaton and Bailey, 1978a). Such information could be made available for a multitude of behavioral procedures (e.g. extinction, time out, differential reinforcement of other behavior), and the initial effects, long term effects, side effects, and rate of change predicted for each procedure provided systematically. Such a "consumer reports" of behavioral procedures would allow for informed choice between alternative procedures for problem behaviors since both the potential benefits and deficits of utilizing specific procedures would be carefully delineated.

Another plausible tack is to design procedures which are not easily degradable or to set up explicit contingencies which make it likely that procedures will be followed carefully. In designing a program to teach young children appropriate street crossing behaviors, Yeaton and Bailey (1978b) intentionally built redundancy into the phases of training so an omission of instructional steps in one phase would not necessarily mean that this aspect of training would be lacking in another phase. Ongoing consultation and corrective feedback by responsible staff members may help to guarantee that the originally trained procedures will continue to be utilized over extended periods of time. The detailed scripting of desired teacher behavior found in the DISTAR materials is another example of an attempt to design instructional materials which are likely to be effective regardless of the trainer, though no contingencies guarantee that the script will be followed.

There does exist research precedent for answering important questions pertaining to the dissemination of treatment packages. Butler (1976) conducted a descriptive study of the success of toilet training procedures utilized by parents after reading Azrin and Foxx's *Toilet Training in Less Than a Day* (1974) and receiving group instructions and feedback. One of the major purposes of the research was "to ascertain the problems reported by parents during and after training." A more controlled investigation of the relative success of several different conditions of implementing the procedures in Azrin and Foxx's book has been conducted by Lutzker and Drake (1976). Yeaton (1979) has demonstrated the decrement in effectiveness when weak treatments are utilized. Anecdotal evidence of some of the potentially adverse side effects resulting when change agents are trained in less than the most desirable manner was also reported.

Dissemination of procedures to consumers was of primary interest in the research of Clark, Greene, Macrae, McNees, Davis, and Risley (1977), and the studies are exemplary in their careful development of an empirically validated treatment package. After a tentative solution had been developed and analysed, these researchers wanted to know the extent to which their solution would be implemented in the absence of the research team and the results of the implementation without quality contol by the experimenters. Such an approach presupposes a marketing orientation, as the satisfaction of the consumer becomes critical to the success of the program in solving the problem (e.g. Risley, 1975). The three studies in this research systematically varied the extent of responsibility of the researchers and the parents involved in teaching appropriate shopping behaviors to their children. The effort is clearly an important first step in the development of a more applied technology of application.

As yet, there has been no clear consensus as to whether innovative treatments should be

implemented exactly as planned, be tailored to the unique situation into which they are placed, or take on an identity in between these two extremes (Calsyn, Tornatzky and Dittmar, 1977; Glaser and Backer, 1977; Larsen and Agarwala-Rogers, 1977). Whichever of these courses is chosen, it is a central thesis of this paper that considerable effort should be expended to preserve the demonstrated effectiveness of these treatments, whether by graphic, social validational, or cost analytic means, as they are disseminated to settings of relevance. For it is the direct assessment and ensured maintenance of effectiveness that may be the most socially significant problem that we, the society of applied behavior analysts, have yet faced.

REFERENCES

Acton, J. P. Economic analysis and the evaluation of medical program. In *Emergency medical services; Research methodology* (DHEW Publication No. (PHS-78-3195). Hyattsville, Maryland: National Center for Health Services Research (1977).

Azrin, N. H. and Foxx, R. M. *Toilet training in less than a day.* New York: Simon and Schuster (1974).

Baer, D. M.(1975) In the beginning, there was a response. In: E. Ramp and G. Semb (Eds.), *Behavior analysis: Areas of research and analysis.* Englewood Cliffs, New Jersey: Prentice Hall.

Baer, D. M. (1977) Perhaps it would be better not to know everything. *J. appl. Behav. Analysis,* **10,** 167–172. (a)

Baer, D. M. (1977) Reviewer's comment: Just because it's reliable doesn't mean that you can use it. J. appl. Behav. Analysis, **10,** 117–119. (b)

Baer, D. M. and Wolf, M. M. (1970) The entry into natural communities of reinforcement. In: R. Ulrich, T. Stachnik, and J. Mabry (Eds.), *Control of human behavior: Volume II.* Glenview, Illinois: Scott Foresman.

Baer, D. M., Wolf, M. M. and Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *J. appl. Behav. Analysis,* **1,** 91–97.

Bailey, J. S. *Handbook of research methods in applied behavior analysis.* New York: Plenum, in press.

Bailey, J. S. Personal communication, October 13, 1978.

Barber, R. M. and Kagey, J. R. (1977) Modification of school attendance for an elementary population. *J. appl. Behav. Analysis,* **10,** 41–48.

Bickman, L. and Zarantonello, M. (1978) The effects of deception and level of obedience on subjects' rating of the Milgram study. *Pers. Soc. Psychol. Bull.* **4,** 81–85.

Boruch, R. F. and Gomez, H. (1977) Sensitivity, bias, and theory in impact evaluations. *Professional Psychology,* **8,** 411–434.

Butler, J. F. (1976) The toilet training success of parents after reading *Toilet training in less than a day. Behav. Therap.* **7,** 185–191.

Calsyn R. J., Tornatzky, L. G. and Dittmar, S. (1977) Incomplete adoption of an innovation: The case of goal attainment scaling. *Evaluation,* **4,** 127–130.

Campbell, D. T. Assessing the impact of planned social change. In: G. M. Lyons (Ed.) *Social research and public policies.* Hanover, N. H., Dartmouth, 1975.

Campbell, D. T. and Fiske, D. W. (1959) Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychol. Bull:* **56,** 81–105.

Campbell, D. T. and Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, (1966).

Chapman, C. and Risley, T. R. (1974) Anti-litter procedures in an urban, high density area. *Journal of app. Behav. Analysis.* **7,** 377–383.

Charters, W. W., Jr. and Jones, J. E. (1973) On the risk of appraising non-events in program evaluation. *Educational Researcher,* **2,** 5–7.

Ciarlo, J. A. (1977) Monitoring and analysis of mental health program outcome data. *Evaluation,* **4,** 109–114.

Clark, H. B., Greene, B. F., Macrae, J. W., McNees, M. P., Davis, J. L. and Risley, T. R. (1977) A parent advice package for family shopping trips: Development and evaluation. J. appl. Behav. Analysis, **10,** 605–624.

Cohen, J. *Statistical power analysis for the behavioral sciences.* New York: Academic Press, 1969.

Cook, T. D. and Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In: M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand McNally, 1976.

Deitz, S. M. (1978) Current status of applied behavior analysis: Science versus technology. *Am. Psychologist,* **33,** 805–814.

DeProspero, A. *A comparison of visual and statistical analyses of intrasubject replication data.* Paper presented at the Association for the Advancement of Behavior Therapy, New York, 1976.

DeProspero, A. and Cohen, S. (1979) Inconsistent visual analyses of intrasubject data. *J. appl. Behav. Analysis,* **12,** 573–579.

Fairweather, G., Sanders, D. and Tornatzky, L. *Creating changes in mental health organizations.* Elmsford, New York: Pergamon Press, 1974.

Fawcett, S. B. and Miller, L. K. (1975) Training public-speaking behavior: An experimental analysis and social validation. *J. appl. Behav. Analysis,* **8,** 125–135.

Fawcett, S. B., Mathews, R. M. and Fletcher, R. K. (1980) Some promising dimensions for behavioral community technology. *J. appl behav. Analysis,* **13,** 505–518.

Feldt, L. S. (1973) What size sample for methods/materials experiments? *J. Educational Measurement,* **10,** 221–226.

Foxx, R. M. and Azrin, N. A. (1972) Restitution: A method of eliminating aggressive-disruptive behavior of retarded and brain damaged patients. *Behav. Res. Ther.,* **10,** 15–27.

Frederiksen, L. W., Jenkins, J. O., Foy, D. W. and Eisler, R. M. (1976) Social-skills training to modify verbal outbursts in adults. *J. appl. Behav. Analysis,* **9,** 117–125.

Freeman, H. E. The present status of evaluation research. In: M. Guttentag (Ed.), *Evaluation studies review annual* (Vol. 2). Beverly Hills: Sage, 1977.

Glaser, E. M. and Backer, T. E. (1977) Innovation redefined: Durability and local adaptation. *Evaluation,* **4,** 131–135.

Glass, G. V., Willson, V. L. and Gottman, J. M. *Design and analysis of time series experiments.* Boulder: Colorado Associated University Press, 1975.

Glenwick, D. and Jason, L. (Eds.) *Behavioral community psychology: Progress and prospects.* New York: Praeger, 1980.

Gori, G. B. and Lynch, C. J. (1978) Towards less hazardous cigarettes: Current advances. *J. Am. med. Assoc.,* **240,** 1255–1259.

Hall, G. E. and Loucks, S. F. (1977) A developmental model for determining whether the treatment is actually implemented. *Am. Educational Res. J.,* **14,** 263–276.

Huff, D. *How to lie with statistics.* New York: W. W. Norton and Co., 1954.

Jacobs, H. *Behavior systems analysis in the development of a community based resource recovery program.* Unpublished doctoral dissertation, Florida State University, 1979.

Jones, R. R., Vaught, R. S. and Weinrott, M. (1977) Time-series analysis in operant research. *J. appl. Behav. Analysis,* **10,** 151–166.

Jones, R. R., Weinrott, M. R. and Vaught, R. S. (1978) Effects of serial dependency on the agreement between visual and statistical inference. *J. appl. Behav. Analysis,* **11,** 277–283.

Kazdin, A. E. *Behavior modification in applied settings.* Homewood, Illinois: Dorsey Press, 1975. (a)

Kazdin, A. E. (1975) Characteristics and trends in applied behavior analysis. *J. appl. Behav. Analysis,* **8,** 332. (b)

Kazdin, A. E. Statistical analyses for single-case experimental designs. In: M. Hersen and D. H. Barlow (Eds.), *Single-case experimental designs: Strategies for studying behavioral change.* Oxford: Pergamon, 1976.

Kazdin, A. E. (1977) Assessing the clinical or applied importance of behavior change through social validation. *Behav. Modification,* **1,** 427–452.

Kazdin, A. E. (1980) Acceptability of alternate treatments for deviant child behavior. *J. appl. Behav. Analysis,* **13,** 259–273.

Kilmann, P. R. and Sotile, W. M. (1976) The marathon encounter group: A review of the outcome literature. *Psycho. Bull.,* **83,** 827–850.

Larsen, J. K. and Agarwala-Rogers, R. (1977) Re-invention of innovative ideas: Modified? Adopted? None of the above? *Evaluation,* **4,** 136–140.

Lave, L. B. and Weber, W. E. (1970) A benefit-cost analysis of auto safety features. *Applied Economics,* **2,** 265–275.

Lesser, G. S. *Children and television: Lessons from Seasame Street.* New York: Random House, 1974.

Levin, H. M. Cost-effectiveness analysis in evaluation research. In: M. Guttentag and E. L. Struening (Eds.), *Handbook of evaluation research* (Vol. 2). Beverly Hills: Sage, 1975.

Lippencott, E. C. *Behavioral nutrition: A public feedback system for the reduction of plate waste in an elementary school.* Unpublished master's thesis, Florida State University, 1978.

Little, A. D., Inc. *Cost-effectiveness in traffic safety;* New York: Frederick A. Praeger, Publishers, 1968.

Lutzker, J. R. and Drake, J. A. *A comparison of trainer-training techniques to produce rapid toilet training in children.* Paper presented at the meeting of the American Psychological Association, Washington, D. C., September, 1976.

McMichael, J. S. and Corey, J. R. (1969) Contingency management in an introductory psychology class produces better learning. *J. appl. Behav. Analysis,* **2,** 79–84.

Michael, J. (1974) Statistical inference for individual organism research: Mixed blessing or curse? *J. appl. Behav. Analysis,* **7,** 647–653.

Milgram, S. (1963) Behavioral study of obedience. *J. Abnormal Social Psychol.,* **67,** 371–378.

Minkin, N., Braukmann, C. J., Minkin, B. L., Timbers, B. J., Fixsen, D. L., Phillips, E. L. and Wolf, M. M. (1976) The social validation and training of conservational skills. *J. appl. Behav. Analysis,* **9,** 127–140.

Nachmias, J. and Steinman, R. M. (1965) An experimental comparison of the method of limits and the double staircase method. *Am. J. Psychol.* **78,** 112–115.

Nagel, S. S. and Neef, M. Determining an optimal level of statistical significance. In: M. Guttentag (Ed.), *Evaluation studies review annual* (Vol. 2). Beverly Hills: Sage, 1977.

Niskanen, W. A., Harberger, A. C., Haveman, R. H., Turvey, R. and Zeckhauser, R. (Eds.). *Benefit-cost and policy analysis, 1972.* Chicago, Aldine, 1973.

O'Brian, C. P., Hamm, K. B., Ray., B. A., Pierce, J. F., Luborsky, L. and Mintz, J. (1972) Group *vs.* individual psychotherapy with schizophrenics. *Arch. gen. Psychiat.,* 1972, **27,** 474–478.

O'Leary, K. D. (1977) Editorial. *J. appl. Behav. Analysis,* **10,** iii–iv.

Parsonson, B. S. and Baer, D. M. The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluting change.* New York; Academic Press, 1978.

Paul, G. L. and Lentz, R. J. *Psychosocial treatment of chronic mental patients.* Cambridge, Mass.: Harvard University Press, 1977.

Phillips, E. L., Phillips, E. A., Wolf, M. M. and Fixsen, D. L. (1973) Achievement place: Development of the elected manager system. *J. appl. Behav. Analysis,* **6,** 541–561.

Porterfield, J. K., Herbert-Jackson, E. and Risley, T. R. (1976) Contingent observation: An effective and acceptable procedure for reducing disruptive behavior of young children in a group setting. *J. appl. Behav. Analysis,* **9,** 55–64.

Prest, A. R. and Turvey, R. (1965) Cost-benefit analysis: A survey. *Economic Journal,* **75,** 683–735.

Quilitch, H. R. (1975) A comparison of three staff-management procedures. *J. appl. Behav. Analysis,* **8,** 59–66.

Risley, T. R. Certify procedures, not people. In: W. S. Wood (Ed.), *Issues in evaluating behavior modification.* Champaign: Research Press, 1975, 159–181.

Ross, H. L., Campbell, D. T. and Glass, G. V. (1970) Determining the social effects of a legal reform: The British "breathalyzer" crackdown of 1967. *Am. Behavior. Scientist,* **13,** 493–509.

Rothenberg, J. Cost-benefit analysis: A methodological exposition. In: M. Guttentag and E. L. Struening (Eds.), *Handbook of evaluation research* (Vol. 2). Beverly Hills: Sage 1975.

Schramm, C. J. (1977) Measuring the return of program costs: Evaluation of a multi-employer alcoholism treatment program. *Am. J. Public Health,* **67,** 50–51.

Schultze, C. L. *The politics and economics of public spending.* Washington, D. C.: The Brookings Institution, 1968.

Sechrest, L. *Estimating size of effects in health research* (Grant application to Department of Health, Education, and Welfare). Unpublished manuscript, Florida State University, 1976.

Sechrest, L. and Redner, R. *Strength and integrity of treatments in evaluation studies.* In: *Evaluation reports:* Washington, D. C., National Criminal Justice Reference Service, 1979.

Sechrest, L. and Yeaton, W. H. Assessing the effectiveness of social programs: Methodological and conceptual issues. In: S. Ball (Ed.), *Assessing and interpreting outcomes.* San Francisco: Jossey-Bass New directions for program evaluation series, 1981, (a).

Sechrest, L. and Yeaton, W. H. Empirical bases for estimating effect size. In: R. F. Boruch, P. M. Wortman, D. S. Cordray and Associates (Eds.). *Reanalyzing program evaluations: Policies and practices for secondary analysis of social and educational programs.* San Francisco: Jossey-Bass, 1981. (b)

Sechrest, L. and Yeaton, W. H. Magnitudes of experimental effects in social science research. *Evaluation Review,* in press.

Shafto, F. and Sulzbacher, S. (1977) Comparing treatment tactics with a hyperactive preschool child: Stimulant medication and programmed teacher intervention. *J. appl. Behav. Analysis,* **10,** 13–20.

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

Smith, M. L. and Glass, G. V. Meta-analysis of psychotherapy outcome studies. *Am. Psychologist,* 1977, **32,** 752–760.

Smith, M. L., Glass, G. V. and Miller, T. I. *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press, 1980.

Stein, T. J. (1975) Some ethical considerations of short-term workshops in the principles and methods of behavior modification. *J. appl. Behav. Analysis,* **8,** 113–115.

Stokes, T. F. and Fawcett, S. B. (1977) Evaluating municipal policy: An analysis of a refuse packaging program. *J. appl. Behavior Analysis,* **10,** 391–398.

Stolz, S. B. (1981) Adoption of innovations from applied behavioral research: "Does anybody care?". *J. appl. Behav. Analysis,* **14,** 491–505.

Valavanis, S. (1958) Traffic safety from an economist's point of view. *The Quarterly Journal of Economics,* **72,** 477–484.

Van Houten, R. (1978) Normative data: A comment. *J. appl. Behav. Analysis,* **11,** 110.

Van Houten, R. (1979) Social validation: The evolution of standards of competency for target behavior. *J. appl Behav. Analysis,* **12,** 581–591.

Walker, H. M. and Hops, H. (1976) Use of normative peer data as a standard for evaluating classroom treatment *effects. J. appl. Behav. Analysis,* **9,** 159–168.

Weisbrod, B. A. Income redistribution effects and benefit-cost analysis. In S. B. Chase, Jr. (Ed.), *Problems in public expenditures analysis.* Washington, D. C.: The Brookings Institution, 1968.

Weisbrod, B. A. Deriving an implicit set of governmental weights for income classes. In: R. Lanyard (Ed.), *Cost-benefit analysis.* Baltimore: Penguin, 1972.

White, M. A. (1975) Natural rates of teacher approval and disapproval in the classroom. *J. appl. Behav. Analysis,* **8,** 367–372.

Willner, A. G., Braukmann, C. J., Kirigin, K. A., Fixsen, D. L., Phillips, E. L. and Wolf, M. M. (1977) The training and validation of youth-preferred social behaviors of child-care personnel. *J. appl. Behav. Analysis,* **10,** 219–230.

Wolf, M. M. (1978) Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *J. appl. Behav. Analysis,* **11,** 203–214.

Yeaton, W. H. *An analysis of a school crossing guard training program: Teaching pedestrian safety to young children.* Unpublished doctoral dissertation, Florida State University, 1979.

Yeaton, W. H. and Bailey, J. S. *Behavioral community psychology: From demonstration to diffusion.* In: J. S. Bailey (Chair), Tactics and techniques in behavioral community psychology. Symposium presented at the meeting of the Midwestern Association of Behavior Analysis, Chicago, 1978. (a)

Yeaton W. H. and Bailey, J. S. (1978) Teaching pedestrian safety skills to young children: An analysis and one-year followup. *J. appl. Behav. Analysis,* **11,** 315–329. (b)

Yeaton, W. H., Green, B. F. and Bailey, J. S. Behavioral community psychology strategies and tactics for teaching community skills to children and adolescents. In: A. E. Kazdin and B. B. Lahey (Eds.), *Advances in clinical child psychology* (Vol. 4). New York: Plenum Press, 1981.

Yeaton, W. H. and Redner, R. Measuring strength and integrity of treatments: Rationale, techniques, and examples. In: R. Conner (Ed.), *Methodological advances in evaluation research.* Beverly Hills: Sage-research progress series in evaluation, 1981.

Yeaton, W. H. and Sechrest, L. (1981) Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *J. Consult. Clin. Psychol.,* **49,** 156–167