JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1987, 20, 313–327 NUMBER 4 (WINTER 1987)

SOME STILL-CURRENT DIMENSIONS OF APPLIED BEHAVIOR ANALYSIS

DONALD M. BAER
MONTROSE M. WOLF
UNIVERSITY OF KANSAS

TODD R. RISLEY
UNIVERSITY OF ALASKA

Twenty years ago, an anthropological note described the current dimensions of applied behavior analysis as it was prescribed and practiced in 1968: It was, or ought to become, applied, behavioral, analytic, technological, conceptual, effective, and capable of appropriately generalized outcomes. A similar anthropological note today finds the same dimensions still prescriptive, and to an increasing extent, descriptive. Several new tactics have become evident, however, some in the realm of conceptual analysis, some in the sociological status of the discipline, and some in its understanding of the necessary systemic nature of any applied discipline that is to operate in the domain of important human behaviors.

DESCRIPTORS: application, dissemination, technology, terminology, history

Twenty years ago, an anthropologist's account of the group calling its culture Applied Behavior Analysis (ABA) had to begin by describing the relevant context (Baer, Wolf, & Risley, 1968): the existence and power of the disciplinary matrix (Kuhn, 1970, p. 175) within which the behavior of individuals was analyzed experimentally. That matrix was itself the characteristic behavior of a more inclusive, older group calling its culture The Experimental Analysis of Behavior (TEAB). The characteristic exemplary strategies (Kuhn, 1970, p. 189) of the TEAB group were their procedural emphases of reinforcement, punishment, and discriminative-stimulus contingencies as behavior-analytic environmental variables, their reliance on single-subject designs as the formats of analysis and proof, and their consistent use of the Skinner box as their arena. Within a decade, the ABA group outnumbered its originally overarching TEAB group, such that the inevitable debates about their actual and desirable conceptual cohesion and separateness took on some sociological urgency (but not very much urgency, especially from nonbehavior analytic points of view. At a conference on behavior analysis, steady-state argument led Nancy Datan to recount an Arab proverb about the nature of conflict: "I must defend my tribe against the world, my family against the tribe, my brothers against my family, and myself against my brothers.") (e.g., Baer, 1981; Deitz, 1978; Michael, 1980; Pierce & Epling, 1980). Even so, the ABA subgroup continued to show nearly the same strategies that characterized the TEAB group, but shifted their application exclusively to what the group called "socially important behaviors" in those behaviors' real-life Skinner boxes.

That shift in strategy required a considerable number of new tactics. At the most basic disciplinary-matrix level, measurement procedures were almost immediately opened to continuous interval-based response measures; the recording not of discrete-response occurrences but instead of intervals during which the response had occurred at least once or was ongoing solved the otherwise unmanageable problem of recording real-life behaviors with difficult-to-define onsets and offsets (Baer, 1986, but cf. Powell, 1984). These measures at first supplemented, and soon almost replaced, the rate-of-response measures characteristic of the parent TEAB group. In addition, seven classes of tactic labels were proposed as stimulus controls for appropriate behavior-analytic conduct in the new world of application (within which behavior-analytic logic is indeed difficult to defend—ironically, the part of the world that likes to call itself "real" usually prefers mentalistic explanations of its own behavior).

The stimulus controls proposed for behavior-analytic conduct in the world of application were the seven key words in a set of injunctions always to be: *applied, behavioral, analytic, technological, conceptual, effective,* and capable of appropriately *generalized* outcomes.

Today, those tactic labels remain functional; they still connote the current dimensions of the work usually called applied behavior analysis. The tactics for which they are stimulus controls have changed to some extent, however. (If they had not changed to some degree in two decades, we might well worry about the viability of their discipline; if they had changed too much, we might well wonder if there was any discipline in their viability. Thus, we would do well to estimate often how properly situated we are between those two extremes.)

Applied

Initially, the meaning of *applied* centered on vague concepts of social problems, social interest, and the immediate importance of the behavior or its functional stimuli to the behaver. Twenty years of experience, especially with what often is called social criticism, have begun to clarify what social problems, interest, and importance are. On the face of it, they are at least behaviors of a person called subject or client that trouble that person; but more often, they are also behaviors of people other than the one called subject or client. Social problems are those behaviors of the subject or client that result in counteraction, sometimes by the client, but more often by nonclients, sufficient to generate something called a solution, or at least a program. (In the world of application, attractive programs that do not solve the problem to which they are ostensibly applied sometimes are valuable even so. At least, they solve the sometimes more aversive problem of doing nothing about that problem. In addition, they very often solve some quite important related problem: They let the client or the counteracting nonclients discuss the problem with a sympathetic friend, they provide those people a platform, or both. Perhaps there is no such thing as a totally ineffective program. But when programs do not solve the target problem, it is typical—and functional—not to measure their ineffectiveness at that, yet it could be illuminating to measure their social validity.)

Thus, social problems are essentially the behaviors of displaying or explaining problems—one's own or someone else's. Problem displays are sometimes large-scale, sometimes small-scale. Perhaps the smallest scale display is seen when one client explains a personal problem to a therapist; the question is whether the client can explain well enough to secure the therapist's attempt at its solution. By contrast, sometimes an entire society can approach nuclear annihilation and technological illiteracy; the question then is whether its media can display and explain that problem effectively enough to secure the political behavior that will generate its government's attempt at solutions, or whether its government will try to solve other, smaller problems, exactly because the small-problem proponents are more effective at using the media, lobbying, and financial campaign support.

It is clear that the therapist's response is usually controlled not simply by the client's promise to pay a fee but also by the therapist's agreement that this problem deserves a solution—an agreement sometimes withheld. Thus, most therapists would consider teaching a self-instructional program aimed at improving a client's dart-throwing skill for social events at a favorite bar, but not at improving the client's rifle-shooting accuracy for a proposed murder. Similarly, the government's decision may (we hope) be controlled not simply by what will accomplish its reelection and the back-up reinforcers pursuant to that, but also by its analysis of its society's survival and prosperity.

The polarities of these two decisions seem to be, respectively, the client's problem display and willingness to pay versus the therapist's values (in other words, the historical and current contingencies controlling the therapist's agreement to program the necessary behavior changes), and the lobbyists' problem displays and willingness to support campaigns versus the government's analysis of societal survival and prosperity (in other words, the historical and current contingencies controlling the government's agreement to program the necessary behavior changes). Those polarities have not changed much in two decades (or in two millenia); what is grimly new in the discipline is the more widespread explicit recognition that all such polarities are themselves behaviors of displayers and displayees; that the behaviors of displaying and explaining problems always exist on continua of their effectiveness for a given displayee; and that whenever one agency displays and explains its problems effectively, its effectiveness can cause some other agency to display that very effectiveness as *its* problem, perhaps more effectively, and so on, *ad infinitum.*

The past two decades have not yielded a better public analysis of effective problem display and explanation (although its deliberate practice is surely one of the world's older professions). At best, they have shown us that we need analyses of (a) displaying and explaining problems so as to gain effective use of the media, (b) controlling the behavior of those other people who can function as decision-makers' constituencies (i.e., lobbying), (c) having or being able to recruit campaign support, and (d) recognizing events called *crises* as the setting events when those repertoires will be most effective. At least those analyses are necessary to understand fully what we most often mean by *applied.* We mean every form of countercontrol typically under the stimulus control of problem displays and explanations. That leaves us with a very large programmatic question: What do we know and what can we learn about effective stimulus control that can be applied in the domain of problem displays? It is clear that some people in our society know a great deal about that. If they know, then we can learn. The crucial behavior may be to establish the priority of that research area as essential to making us a truly applied discipline; clearly, the last two decades have prompted that priority with increasing urgency.

Behavioral

One mark of the success of applied behavior analysis in the last two decades is that its practitioners, researchers, and theorists have encountered so many invitations to become something other than behavioral, usually in the form of becoming something "more" than behavioral. In particular, their occasional mainstreaming with behavior therapy, education, developmental psychology, psycholinguistics, and sociobiology has given them the chance to entertain constructs of anxiety, attention, intelligence, disabilities, spontaneity, readiness, critical periods, innate releasers, storage and retrieval mechanisms, schemata, and the like. Some behavior analysts did entertain one or more of those constructs enough to be no longer behavioral; others were simply entertained by those constructs. The most fruitful task, however, is to recognize that each of those labels (and many others like them) often represents some behavioral reality not yet analyzed as such. The point is that these behavioral realities are not likely to be analyzed as such within their parent disciplines, and thus never will become truly applicable there, yet might well be analyzed behavior-analytically, perhaps with great profit to us and those disciplines, and thus to our roles within those disciplines.

Doing so will not jeopardize our ability to discriminate a behavioral discipline from a nonbehavioral discipline: The various professional behavior patterns that constitute a behavioral discipline, thoroughly described and analyzed by Zuriff (1985), can always be discriminated from the considerably more various patterns that constitute nonbehavioral disciplines, even if no one were any longer to display those behavior patterns. (In other words, Zuriff's analysis is essentially philosophical rather than anthropological.) However, it seems clear that behaviorism will be a small-minority approach, at least for the foreseeable future of this culture. Indeed, behavioral textbooks explaining the Premack principle might include in their lists of cultural reinforcers access to the use and consumption of inner, mentalistic explanations for behavior. Perhaps behavior-analytic language is the key to that. The past 20 years have shown us again and again that our audiences respond very negatively to our systematic explanations of our programs and their underlying assumptions, yet very positively to the total spectacle of our programs—their procedures and their results—as long as they are left "unexplained" by us.

Hineline (1980) has begun the analysis of how our systematic language affects our audiences, and how they use their own unsystematic language to explain behavior. Sometimes, for example, certain contexts actually do evoke attributions of behavior to environmental causes, yet even that kind of attribution and its contextual control can themselves be attributed to internal "personality" causes, and in a language culture like ours, they usually are (see Hineline, 1980, p. 84). Perhaps applied behavior analysis should consider much more carefully and much more explicitly the language options that might maximize its effectiveness in its culture: (a) find ways to teach its culture to talk behavior-analytically (or at least to value behavior-analytic talk); (b) develop nonbehavior-analytic talk for public display, and see if that talk will prove as useful for research and analysis as present behavior-analytic talk, or whether two languages must be maintained; or (c) let it be (we represent approximately 2% of American psychology, and we are currently stable at that level).

Some of the success of applied behavior analysis has led to its trial in office-practice contexts. In those contexts, the direct observation of behavior often seems impractical, and practitioners resort to more suspect forms of observation, for example, self-reports or ratings by participant-observers, both often in the form of answers to questionnaires, inventories, checklists, interviews, focused diaries, and the like. With such measures, it is considered safer to use many of them at the same time (see *Behavioral Assessment,* 1979–). The thesis that one behavior can be a measure of another behavior seems behavioral on its face; on analysis, it seems behavioral but extraordinarily risky, depending heavily on the choice of the "other" behavior.

Twenty years of practice have given applied behavior analysis a nearly standard measurement method: the direct observation and recording of a subject's target behaviors by an observer under the stimulus control of a written behavior code. Obviously, that is the measurement of some behavior of one person by some other behavior of another person. The strength of this particular method is the modifiability of the observer's behavior by careful, direct training, and the accessibility of the observer's behavior to direct and frequent reliability assessments. In particular, when those reliability assessments pair the observer with the code-writer, they accomplish the essential validity of any observation-based behavioral analysis: They allow the empirical revision of the code and thus of the stimulus control that it exerts over the observer's observing and recording behavior, until it satisfies the code-writer. That revision is accomplished by rewriting the code and retraining the observer's response to it until the observer's recordings of target behavior agree closely with those of the code-writer observing the same sample of the subject's behavior. Thus, the code-writer *controls* the observing and recording behavior of the observer, and in principle can assess and refine that control as often as the problem may require. In that the code-writer is (or ought to be) the person who finds the subject's behavior to be a problem (or is the surrogate of that person), then satisfying the code-writer that all and only the correct behaviors are being recorded is the *only* approach to valid measurement that makes complete systematic sense: Valid measurement is measurement of that behavior that has caused the problem-presenter to present it (cf. Baer, 1986). Clearly, this argument does not change if the observer is replaced by a recording instrument. This is a strong argument against the use of standardized codes: It is unlikely that a standardized code written in complaint of someone else's behavior can satisfy the present complainer as well as the code that this complainer would write about this specific complainee.

By contrast, it is risky to assume that the subject's self-report or a participant-observer's rating of the subject's target behavior would show a similar reliability with its direct observation by an observer under the code-writer's control. The observer's behavior can be controlled in a well-understood manner; the subject's self-reports and the participant-observer's ratings usually are uncontrolled by the practitioner-researcher. In principle, the subject's self-reports and the participant-observers' ratings might be controlled in the same way as are a standard observer's observation and recording behaviors, but we know relatively little about doing so, and although we often can maintain nearly exclusive control of the observer's relevant behavior, we rarely can even approach that exclusivity with the subject's or a participant-observer's behavior.

Of course, self-reports and participant-observers' ratings might be studied in their own right as behaviors for analysis, rather than as substitutes for the direct observation of the target behavior. Their analysis would almost certainly yield interesting knowledge of the large world of verbal behavior and the small world of professional ritual, but apart from that, it would not often seem to have applied significance, other than to document the already strongly suspected invalidity of such behaviors as substitutes for the target behavior. However, within that small world of professional ritual, it is worth noting that the use of such measures—often called psychometrics in the social sciences—has a certain social validity, especially for research-grant applications: Some role of conventional psychometrics in a research proposal increases the probability of that proposal being approved and funded when the proposal's reviewers are not behavior-analytic (which is almost always). It is true that any choice of Psychometric$_1$ versus Psychometric$_2$ will inevitably attract at least some reviewers' criticisms, but at least it will be criticism within the context of approval for playing the correct game. That might be considered applied significance.

Thus, applied behavior analysis most often still is, and most often always should be, the study of an observer's behavior that has been brought under the tight control of the subject's behavior. Sometimes, that is exactly what is meant by behavioral assessment. More often, it is either the study of how subjects talk about their own behavior or how other people talk about the subject's behavior, a kind of talk that usually is under complex, varied, and largely unknown control, only one component of which *may* be the subject's target behavior (what loosely is called the truth).

Sometimes, though (and increasingly in the past two decades), behavioral assessment has used those forms of psychometrics that are best described as *samples* of the target repertoire, notably IQ and achievement tests. The problems with such tests are much the same as with self-reports and participant-observers' ratings: We rarely know if the testing context controls those behaviors differently than they are controlled in everyday life, and we rarely know if those test samples are representative samples of the desired repertoire. The only way to know those facts with any certainty is again to resort to direct observation, but these tests represent (or fail to represent) repertoires that often are too large to allow practical direct observation. Thus they are often used as the only practicable alternative, despite their uncertainties (and sometimes they are used because they still command great social validity in this society).

That tactic is not a novel one in applied behavior analysis: In the analysis of accidents, for example, we can hardly deal with accident behaviors directly, because they are too infrequent, so we change the much more frequent behaviors that we suppose are precursors to accidents—we analyze not accidents, but risk-taking. Similarly, in the analysis of delinquency, we can hardly change delinquent acts directly, again because they are infrequent and also because they are systematically done in relative secrecy, so again we change not them but what we suppose their precursors are in various arenas of social control. If the guesses implicit in those areas of research do not disqualify them as examples of applied behavior analysis, then the analogous guesses implicit in the use of, say, achievement tests need not automatically disqualify them, either.

The applied question most often may be whether the uncertainties inherent in resorting to such measures are preferable to the status quo of knowledge in each problem area, and the answer, like most answers, will probably be under contextual control—sometimes uncertainty is preferable to status quo, sometimes it isn't.

Thus the term "behavioral assessment," a new category event of the past two decades, sometimes describes very pragmatic tactics and sometimes only the least valid measurement tactics of the very old pseudobehavioral disciplines against which behavior analysis rebelled. Clearly, its tactics can include exceptionally elegant and sophisticated concepts, techniques, and measures of reliability (see Cronbach, Glaser, Nanda, & Rajaratnam, 1972), which sometimes are applicable to direct observation (Hartmann, 1977); but when those tactics measure what we wish to analyze is problematic in both analytic and pragmatic ways. Ultimately, knowing when they do and when they do not will require very difficult studies based on direct observation.

Analytic and Conceptual

Twenty years ago, *analytic* meant a convincing experimental design, and *conceptual* meant relevance to a comprehensive theory about behavior. The two topics could be and often were discussed separately. Since then, it has become increasingly aversive to maintain that separation. Now, applied behavior analysis is more often considered an analytic discipline only when it demonstrates convincingly how to make specified behavior changes *and* when its behavior-change methods make systematic, conceptual sense. In the past 20 years, we have sometimes demonstrated convincingly that we had changed behavior as specified, but by methods that did not make systematic, conceptual sense—it was not clear *why* those methods had worked. Such cases let us see that we were sometimes convincingly applied and behavioral, yet even so, not sufficiently analytic. Similarly, we have sometimes changed behavior without even a convincing demonstration of how we did that, and so did not know if our methods made systematic, conceptual sense because we did not know clearly what the responsible methods were; those cases let us see how not to be a discipline, let alone an applied, behavioral, or analytic one.

Now, the theory that defines systematic, conceptual sense for us is pushed not only to be about behavior, but also about the behavior of changing behavior: More often now, we can see ourselves as the subjects of someone else, not just as Experimenter (recall the discussion of countercontrol under *Applied*). This fits well with the steadily emerging contextualism apparent in unapplied behavior analysis. A proper appreciation of context always implies that we are not merely studying or managing it, but also are part of it and therefore are being managed by it, even down to our studying and managing of it.

The emerging appreciation of context as the *setting events* that had better be understood and managed in truly effective application flows easily from Kantor's field approach to the study of behavior (Morris, 1982). But it also flows just as easily from our recently expanding knowledge and management of stimulus control and conditional stimulus control (see Sidman, 1986; and the special-issue Volume 6 of *Analysis and Intervention in Developmental Disabilities,* 1986). That development suggests strongly that we will rarely find an instance of stimulus control not modified drastically by some (and perhaps many) conditional stimulus controls. The relevance of that thesis to application is urgent: It begins the analysis of the generality of any intervention's effectiveness, in that it urges us to seek the contextual conditions under which the intervention has maximal and minimal effectiveness.

Thus, the first applied lesson of contextualism is that there will always be such conditions; the second is that many of them must be clarified as stimulus and response events, because that is rarely self-evident (cf. Wahler and Fox's [1982] discussion of "insularity" as a limiting condition in parent training); the third, most difficult yet most pragmatic, is that clarifying contextual controls is not enough: If we want widely effective interventions, we will have to manage these contextual controls; rather than stopping with the assessment of their roles as limiting factors, we will have to learn how to program around them or program them away.

Contextualism also implies a certain class of experimental designs. The simplest contextual statements are of the form, Behavior B is controlled differently by Variable V in Context 1 than in Context 2. To see that reliably, we need experimental control of at least two levels of Variable V, say V1 and V2; and we need experimental control of at least two contexts of interest, say Context X and Context Y. Given that, we need to see how

Behavior B relates to V1 and V2 in Context X, and we need to see if that control is reliable. Then we need to see how Behavior B relates to V1 and V2 in Context Y, and we need to see if that control is reliable. Finally, we need to see both of those relationships (how B relates to V1 and V2 in Context X, and how B relates to V1 and V2 in Context Y) again and again, so that we see whether the difference that Contexts X and Y make in how V1 and V2 control B is a reliable difference. The simplest reversal designs would look something like the following two, where CX and CY are Contexts X and Y (see diagram below). Both of these are minimal designs for the problem, yet each contains 16 conditions in which to examine the ongoing baseline of the behavior under study. The pace of the design had better be a rather fast one, suggesting that variant of the reversal design often called the multielement design (e.g., Ulman & Sulzer-Azaroff, 1975). Designs like these can be found in the literature of the field, but not often. To the extent that applied behavior analysis will analyze rather than assess the generality of its interventions, these designs and others capable of the same kind of demonstration will prove essential.

The last 20 years have seen considerable development of research designs. At the outset, it was sufficient to label only the reversal and multiple baseline designs: the examination of one behavior in repeated experimental conditions, and the examination of many behaviors, sometimes with some in one experimental condition while others are in a different experimental condition. These are the two fundamental analysis strategies, of course; their logic is seen in the multiple, mixed, and concurrent schedules that had so often served as experimental designs in TEAB. But in that world, schedules had names, yet designs did not: Researchers simply arranged those conditions (often, schedules) necessary to answer their experimental questions, refined the conditions as analytically as their knowledge of potential important confounding variables allowed, and did all that as often as conviction required. The value of their designs lay not in any category names that might be imposed on them but in the relation between the conditions that they had arranged and the question proposed.

Now, we have named so many designs that textbooks devoted to their taxonomy and their "rules" have emerged. The strategy underlying that development was probably like the one underlying the seven self-conscious guides to behavior analytic conduct posed in 1968 (applied, behavioral, analytic, etc.): In application, good design would often prove difficult to accomplish or maintain, and graduate training in application might not often plumb the depths of the topic; codification might help overcome those difficulties. The questions now are whether in fact the codification of research design into types and rules did help that purpose; if so, to what extent; and finally, whether that extent is worth the cost. The cost may be primarily that applied researchers increasingly transform questions to fit the known designs and their rules, rather than constructing a design that answers the original question. It might prove valuable to the field to recall its original designs and their logic—a good design is one that answers the question convincingly, and as such needs to be constructed in reaction to the question and then tested through argument in that context (sometimes called "thinking through"), rather than imitated from a textbook. For example, one convention paper evaluated a program training youths to fill out employment applications more effectively. The researchers asked several employers to read a sequence of applications, each written by

CX	CY	CX	CY
V1 V2 V1 V2	V1 V2 V1 V2	V1 V2 V1 V2	V1 V2 V1 V2
VI	V2	V1	V2
CX CY CX CY	CX CY CX CY	CX CY CX CY	CX CY CX CY

a different trainee, beginning with some written before training and ending with some written after training; the change from pre- to posttraining applications occurred at different points in each employer's sequence, in an apparent multiple baseline design across employers. The design was such that all applications, pre- and posttraining alike, were read. Almost without exception, the employers said "No" to trainee applications written before training and "Yes" to trainee applications written after training. This design is alluded to in an article by Mathews and Fawcett (1984), but is not described there in detail because of editorial insistence. Many in the convention audience, perhaps like the editor, ignored the fact that this design showed clearly that the training program was exceptionally effective, and argued instead that it was not a "proper" multiple baseline design. Perhaps the important point is that convincing designs should be more important than "proper" designs.

Technological

Twenty years ago, it was urgent to recommend that a new field aspiring to both effective application and stature as a science be both procedural and explicit about it. The point was to avoid the situation of so many clinical, management, and administrative disciplines in which, once discussions of theory and goals had ended, the procedures to be applied were specified no better than "work with." For the most part, that has happened; journal articles and textbooks do offer a complete list of their operative procedures, such that a reader has a fair chance of replicating the application with the same results. Indeed, collections of procedures have begun to emerge, such that readers now may choose among alternative procedures aimed at the same goal.

Still, three points deserve comment:

1. Some procedures, such as praise or incidental teaching, often are varied in what the researcher considers to be a desirably natural manner from occasion to occasion. Those topographies and their sequences rarely are specified; to do so in advance might often be considered unnatural, and to do so retrospectively (e.g., from a videotape) would be expensive for publishers and boring for readers. The underlying assumption is of course that those variations make no difference to the outcome. That assumption is rarely if ever tested empirically. It would be good for the discipline if a review 20 years from now could state that the assumption had proved correct, or that it had been found incorrect so often that current practice had remedied the problem, despite the expense. (Readers' boredom with such detail would have dissolved in the discovery that these variations could indeed make a difference in outcome.)

2. In application, those procedures carried out by people (which are most of the procedures of applied behavior analysis) usually are observed and recorded, just as are the subject's behaviors. This documents the extent to which the specified procedures are performed, and also describes any unspecified procedures that may occur. The process is probably reactive in many applications, creating greater adherence to specified procedures than might be secured otherwise. But these data are rarely presented outside of the group conducting the application, again probably because of publishers' expense and readers' presumed boredom: When such data show that the specified procedures are being carried out well enough, there is no problem; and when they show the opposite, the application usually stops until better adherence to procedure is obtained, whereupon there is again no problem. This argument is probably defensible on a cost-benefit basis, but it would be better for the discipline if its review 20 years from now could state that the relevant debate had occurred publicly. That debate is essentially a matter for journal and textbook editors and reviewers: They call for such data or fail to; they publish such data when supplied or recommend against doing so. Thus, they might well use one of their future journal symposia to consider this issue, which is mainly a matter of journal policy.

3. Dissemination is a practice much older than applied behavior analysis, but, in the realm of behavior, it is usually much less technological than applied behavior analysis. Even so, its literature and its practitioners debate (without resolution) an es-

sentially technological issue: When a program is disseminated, should its disseminators require that its procedures be followed faithfully, no matter where or when the program is used? Or should its users be allowed, and even encouraged, to modify those procedures to fit their local situations and contingencies? (We might first ask, functionally, when we have that choice. That is, when is the control requisite to maintain fidelity to original procedures available to us, and when not?) Fidelity to original procedures is recommended because those procedures have been studied and are known to be effective; their variations and alternatives usually have not been studied, so nothing can be said about their effectiveness. On the other hand, flexibility in application is recommended on the premise that the entire program will become aversive to people who cannot modify it to suit their situation and their contingencies, and if a program is not used, it cannot be effective.

These are both technological arguments; interestingly, contextualism, experience, and common sense seem to agree that each is likely to be correct in certain contexts but not in others. The empirical investigation of those controlling contexts obviously is crucial to future large-scale dissemination (which is certainly the essence of *applied*), as is the investigation of when we even have that choice. That research has largely not been done; presumably, now that a discipline as technological as applied behavior analysis has entered the domain of dissemination, it is more likely to be done, albeit expensively. The appropriate strategy was recommended by Sidman (1960) almost 30 years ago, in a different but relevant context: One criterion of important science is to explore the controlling conditions of any behavioral phenomenon. What is the range of variation of a program's procedures that still allows sufficient effectiveness? If it is large enough, flexible application can be encouraged, and the program's survival in diverse settings may well be enhanced. If it is narrow, fidelity will be required, or what survives will not be effective.

It will be interesting to see if a review of the discipline 20 years from now will be able to summarize some facts about those processes, or will instead have to report that applied behavior analysis is still entering its large-scale applications very much at risk for failure.

Capable of Appropriately Generalized Outcomes

Twenty years ago, the ability of the discipline to produce appropriately generalized outcomes was seen as crucial to its survival: An applied discipline that had to make every topographical variant of its desired behavior changes, and had to attach each of them to every appropriate stimulus control, across time, was intrinsically impractical. Today, the problem is still crucial, but now to the maximal effectiveness rather than the survival of the discipline. In the past 20 years, we have changed behavior as specified *and* shown experimental control of its appropriate generalization just often enough to make clear that the discipline is capable of such outcomes. What remains is the much more reassuring (and much larger) task of exploring the conditions that control appropriate generalization (i.e., appropriate stimulus control).

Fortunately, the problem is usually seen now as one that probably can be solved by suitable programming, rather than by good luck; thus, a good deal of research has systematically examined ways to teach from the outset so that appropriately generalized outcomes are established. (Yet a remarkable number of studies do not compare their generalization-facilitative teaching to any alternative teaching of the same target behavior that does not facilitate its generalization, and so we actually learn nothing about the problem from such studies.) The problem is far from solved; we still have no system for matching the most suitable generalization-promotion method to the behavior change at hand, and no certainty that there is such a system to be found. Our categorizations of generalization-promotion techniques are clearly nonanalytic; they have been proposed (see Stokes & Baer, 1977) in the same way that current dimensions (applied, behavioral, analytic, etc.) have been proposed—on the assumption that codification will evoke more of the necessary professional behavior, especially research (Baer, 1982). That assumption probably cannot be tested empirically: We can hardly conduct an experiment that compares our discipline's progress toward thorough control of generalization, with and without such codifications. Thus, there remains the obligation of continuing debate (see Johnston, 1979).

Effective

The hallmark of any applied discipline ought to be effectiveness; the case is no different for applied behavior analysis. However, in the realm of behavior change, the hallmark of effectiveness can be subtle: Sometimes, it seems to be simply the degree to which the target behavior has been changed; much more often, it is the degree to which something other than the target behavior has been changed, and that something other almost invariably is someone's countercontrol against the original behavior (see the earlier discussion of *Applied*). Thus, for example, if we look closely, we may find that in some cases, changing a student's grades from Fs to Cs satisfies the student, the student's family, and that segment of their society that will eventually read those grades and react to them—if the grades are Fs, these agents will see that as a problem; if the grades are Cs, they will not. But in some other cases, we may find that a student's grades must be changed from Bs to As before that student, that student's family, and that segment of their society that will eventually read those grades and react to them will stop reacting to them as a problem. The marker variable distinguishing these two cases may often seem to be social class, but that is neither analytic (see Baer, 1984, pp. 547–551) nor relevant to the point, which is that changing grades is not effective per se; stopping and avoiding the relevant references to these grades as a problem is the true criterion of our intervention's effectiveness.

Almost every successful study of behavior change ought to routinely present two outcomes—a measure of the changed target behaviors, of course, and a measure of the problem displays and explanations that have stopped or diminished in consequence. Yet very few studies do that. Perhaps their researchers assume that *they* are the only relevant problem-detectors or problem-detector surrogates. Indeed, that may sometimes be true, but it had better be both defensible and explicitly defended or it becomes arrogance (which may not further the social status of the discipline if it is widely noticed as such). On the other hand, the absence of that second measure may represent a crucial weakness in our current effectiveness. We may have taught many social skills without examining whether they actually furthered the subject's social life; many courtesy skills without examining whether anyone actually noticed or cared; many safety skills without examining whether the subject was actually safer thereafter; many language skills without measuring whether the subject actually used them to interact differently than before; many on-task skills without measuring the actual value of those tasks; and, in general, many survival skills without examining the subject's actual subsequent survival. Some of those measures will be controversial to define and expensive to collect, of course; but it may be true that the discipline has developed to the point at which they become crucial. (Children usually become more expensive as they grow.)

Perhaps this practice will become more widespread in the discipline as the calculation of cost–benefit ratios increases from its present near-zero rate—if we take the "benefit" side of the ratio seriously, rather than assume that the behavior change itself is the benefit. Cost–benefit ratios, on the face of it, are the essence of effectiveness, and ought to be routine in any applied discipline (e.g., Hill et al., 1987). They have proven problematic in this one, perhaps partly because the discipline is still so much in its research-trials phase, partly because behavioral benefits are not as clearly defined as most business benefits, and partly because the concepts and techniques of cost–benefit calculation are not yet clearly established themselves.

Fortunately, at least one second measure of effectiveness is beginning to become routine: social validity (Kazdin, 1977; Wolf, 1978), which is the extent to which all the consumers of an intervention like it (i.e., like its goals, targets, effects, procedures, and personnel). The point of social-validity measures is to predict (and thus avoid) rejection of an intervention, especially when it is disseminated (which, because of its large scale, may prove less tolerable to consumers than the initial small-scale research trials). If an intervention is socially invalid, it can hardly be effective, even if it changes its target behaviors thoroughly and with an otherwise excellent cost–benefit ratio; social validity is not sufficient for effectiveness but is necessary to effectiveness.

Unfortunately, social validity is sometimes assessed at present in very rudimentary ways that may too often find social validity where it does not actually operate. Perhaps the problem is that researchers are in the context of hoping for social validity, which is subtly different from and much more dangerous than the context of searching for any sources of social invalidity. In that the discipline is now moving into large-scale dissemination, valid social-validity assessments will soon become crucial to survival; yet this aspect of our measurement technique has seen very little inquiry and development.

Perhaps a review 20 years from now will report a great deal of progress in that dimension of effectiveness. If so, the problem will not have proved to be simple. For example, some measures of social validity are deliberately and pragmatically biased toward positive results, not to deceive their users but to prevent alarm in their consumers (e.g., governing boards) while at the same time alerting their users (the researcher-appliers) to detect and remedy the problems that must underlie scores that usually are 7 but now are 5 on a 7-point scale. Furthermore, it is entirely possible that even quite invalid queries into social validity are better than no queries at all: Giving consumers any opportunity to express complaints and discontents that otherwise would go unnoticed may save at least some programs from fatal backlashes, at least if the offended consumer is moved enough by simply the existence of the otherwise inadequate social-validity assessment form to write in its margins or talk to the appliers.

Perhaps equally significant is the recent development of assessment techniques that inquire about consumers' goals before the program is designed (Fawcett, Seekins, Whang, Muiu, & Suarez de Balcazar, 1982; Schriner & Fawcett, in press), so that the program has a chance to achieve all of those goals, thereby going far to guarantee validly high social validity when that dimension is eventually assessed. This technology, if pursued intensively enough, may become part of the pragmatic analysis of social validity, especially because it finds common themes emerging from its inquiries about the goals of different sets of consumers in what seem to be quite different problem situations (cf. Seekins, Mathews, Fawcett, & Jones, in press); thus the analysis of its validity may be one of the best priority targets for future research.

Perhaps the clearest measure of our discipline's effectiveness is the increasing number of ineffective applications that we have tried in recent years. By good judgment or good luck, we began with dramatic, troublesome, yet nevertheless crucially delimited cases, and our effectiveness with them strongly reinforced our disciplinary behaviors. Had it been otherwise, we might not be the recognized applied discipline that we are today. But having done that, we are of course moving on to a different class of problems, partly because those problems are there, partly because they are exceptionally important, and partly because we are still a research-based applied discipline, and because research ought not to be too repetitive, then, to the extent that we have done (or at least sampled) everything else, these problems are what is left to do.

But the problems of today are not as delimited as those of our beginnings. They are called lifestyles in recognition of their systemic nature. The behavior classes called delinquency, substance abuse, safety, exercise, and diet, for example, represent complex classes of topographies serving complex functions involving many agents of reinforcement/punishment and stimulus control, all of whom interact to constitute and maintain the system as such. Thus, entry at just one point of such systems is likely to yield only limited, short-term behavior changes before the inevitable countercontrol restores the prior status of the system, with us either frozen out or co-opted ineffectively within (see Wolf, Braukmann, & Ramp, 1987). The first remedy is recognition: The concept of systems analysis is now an important component of our effectiveness, and research that will show us how to do that better will prove exceptionally useful. The second remedy, following whatever analysis the first currently allows, is system-wide intervention: Thus, for example, obese children are dealt with not as simple therapist–client interactions, but within their life systems—at least within their families (Brownell, Kelman, & Stunkard, 1983) and better yet, within their families and their school systems (Brownell & Kaye, 1982). The third remedy may be the discrimination of those problems in which a single, short-duration intervention can be effective from those invariably systemic problems in which chronic presence will be required to maintain the effective intervention. Just as medicine recognizes that appendicitis needs only one intervention but that diabetes needs life-long treatment and educates its consumers to the inevitability of that and its costs, applied behavior analysis had better begin its validation and use of the same two categories.

Perhaps the most important remedy of all, however, will be to establish the proper context in which to respond to failures. The last 20 years have produced an increasing rate of them; the next 20 almost surely will see that rate continue and, very likely, increase even more. That fact and that probability have already been interpreted as an inadequacy of behavior-analytic principles. For example, Reppucci and Saunders (1974) responded to one of their failures in a delinquency institution with a broadly generalized principle:

> Finally, there is an issue the resolution of which will have enormous consequences for behavior modification as we know and apply it today. The issue inheres in the fact [sic] that the principles of behavior modification are insufficient and often inappropriate for understanding natural settings—their structure, goals, tradition, and intersetting linkages. (p. 569)

Its publication in the *American Psychologist* of course presented this new "fact" to potentially every APA member (most of whom would not know that it is the only kind of evaluation of behavior modification that their association's journal ever prints, and who might not ask how a "fact" like that can be established through one failure to install and maintain a program in a single institution).

It is worth asking first if technological failure is the same as theoretical failure. Quite likely, technological failure is an expected and indeed important event in the progress of *any* applied field, even those whose underlying theory is thoroughly valid. Thus, the step from the physics laboratory to engineering has been, and will continue to be, marked by occasional jammed elevators, fallen bridges, crashed airplanes, and exploded space shuttles. The engineers know that; they abandon only their designs, not their theories, with each such event. Petroski (1985), for example, sums up their history as follows:

> I believe that the concept of failure—mechanical and structural failure in the context of this discusson—is central to understanding engineering, for engineering design has as its first and foremost objective the obviation of failure. Thus the colossal disasters that do occur are ultimately failures of design, but the lessons learned from those disasters can do more to advance engineering knowledge than all the successful machines and structures in the world. Indeed, failures appear to be inevitable in the wake of prolonged success, which encourages lower margins of safety. Failures in turn lead to greater safety margins and, hence, new periods of success. To understand what engineering is and what engineers do is to understand how failures can happen and how they contribute more than successes to advance technology. (p. xii)

The same point is inherent in an understanding of medical progress—every death is, in a sense, a failure in our current designs for health maintenance, just as every fallen bridge is a failure in our current designs for balancing load against strength.

Applied behavior analysis must deal with phenomena at least as complex as loaded bridges and stressed physiologies, and perhaps, considering the domains of variables relevant to behavior, more complex. Then it will proceed at first with as many, or more, flawed designs as those fields have; but it will profit from those failures, as they have, and it will require time and repetition to do so, as they have.

How do we know that any given failure reflects bad design rather than inadequate principle? We never know that; but we can search for bad design immediately after every failure, and if we find it, that will give us something to try in the next application much more readily than will despair over our principles. For example, Reppucci and Saunders played the role only of outside consultants in their failure. Indeed, it is the *principles* of behavior analysis (as well as considerable experience) that suggest little potential for changing the behavior of overworked, underinterested staff and administrators with the few contingencies usually available to consultants (unless a severe crisis is ongoing). Liberman (1980), in response to a number of similar failures, has suggested not a new principle but merely a stronger design—that some of us combine research with administration:

> We cannot count on administrators' need for accountability and program evaluation to serve as "coattails" for our behavioral programs. . . . If we want our work to live beyond a library bookshelf, we will have to jump into the political mainstream and get our feet wet as administrator-researchers. (pp. 370–371)

If we survey those behavioral programs that have maintained themselves over impressive spans of time, we as often find the pattern Liberman recommends as we find impressive spans of time: Liberman himself at the Oxnard Mental Health Center, McClannahan and Krantz at the Princeton Child Development Institute, Cataldo at the Johns Hopkins' Kennedy Institute, and Christian at the May Institute, for examples. These cases are not proofs of anything, nor intended to be; they are simply worth considering as designs that might attract a proof and might yield a profit for our discipline if they did.

The Teaching-Family model is another example of a somewhat different and apparently durable design, one as old as this journal. It created 12 regional training centers to mediate 215 replications of the original Achievement Place delinquency program, and wrote not journal articles but plain-English training manuals for their use. The originators of that program also met Liberman's design prescription; they added to their research role those of administering the component programs and fighting their political battles, and of securing consistent enough grant support for the necessary 10 years of trial-and-failure-and-next-trial-and-success research necessary to understand and implement the essential quality control, staff training, and other support systems required for survival and dissemination. Indeed, even 20 years have not seen the completion of that research and development program, but its failures are now rare, despite greatly expanded opportunities (Wolf et al., 1987).

The point is that failures teach; the Teaching-Family model grew out of the Teaching-Failure model. Surely our journals should begin to publish not only our field's successes but also those of its failures done well enough to let us see the possibility of better designs than theirs.

In summary, effectiveness for the future will probably be built primarily on system-wide interventions and high-quality failures, as we continue to bring theory to the point of designs that solve problems. But it should be current theory that is built on, not some replacement of it—current theory has worked far too well to be abandoned in the face of what are more parsimoniously seen as technological rather than theoretical failures. Clearly, increasing our effectiveness will not be easy, and it will not happen quickly. We should expect a long period of difficult, expensive, repetitive, and sometimes ineffective research into these applications, and we should enter that research with our best social skills, because we shall require the cooperation of unusually many people, often in unusually exposed positions. However, even with relatively little reaction-to-failure work behind us, it seems clear that we can do it.

It seems clear that we can do what remains to be done. That we can is probably our most fundamental, most important, and most enduring dimension; that we will is simply logical.

REFERENCES

Analysis and Intervention in Developmental Disabilities. (1986). Special issue, **6.**

Baer, D. M. (1981). A flight of behavior analysis. *The Behavior Analyst,* **4,** 85–91.

Baer, D. M. (1982). The role of current pragmatics in the future analysis of generalization technology. In R. B. Stuart (Ed.), *Adherance, compliance, and generalization in behavioral medicine.* New York: Brunner/Mazel.

Baer, D. M. (1984). Future directions? Or, is it useful to ask, "Where did we go wrong?" before we go? In R. A. Polster & R. F. Dangel (Eds.), *Behavioral parent training: Where it came from and where it's at.* New York: Guilford Press.

Baer, D. M. (1986). In application, frequency is not the only estimate of the probability of behavior units. In M. D. Zeiler & T. Thompson (Eds.), *Analysis and integration of behavioral units* (pp. 117–136). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Behavioral Assessment. (1979–). New York: Pergamon Press.

Brownell, K. D., & Kaye, F. S. (1982). A school-based behavior modification, nutrition education, and physical activity program for obese children. American Journal of Clinical Nutrition, **35,** 277–283.

Brownell, K. D., Kelman, J. H., & Stunkard, A. J. (1983). Treatment of obese children with and without their mothers. Changes in weight and blood pressure. *Pediatrics,* **71,** 515–523.

Cronbach, L. J., Glaser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

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

Fawcett, S. B., Seekins, T., Whang, P. L., Muiu, C., & Suarez de Balcazar, Y. (1982). Involving consumers in decision-making. *Social Policy,* **13,** 36–41.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis,* **10,** 103–116.

Hill, M. L., Banks, P. D., Handrich, R. R., Wehman, P. H., Hill, J. W., & Shafer, M. S. (1987). Benefit-cost analysis of supported competitive employment for persons with mental retardation. *Research in Developmental Disabilities,* **8,** 71–89.

Hineline, P. N. (1980). The language of behavior analysis: Its community, its functions, and its limitations. *Behaviorism,* **8,** 67–86.

Johnston, J. M. (1979). On the relation between generalization and generality. *The Behavior Analyst,* **2,** 1–6.

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

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd Ed.). Chicago: University of Chicago Press.

Liberman, R. P. (1980). Review of: *Psychosocial treatment for chronic mental patients* by Gordon L. Paul and Robert J. Lentz. *Journal of Applied Behavior Analysis,* **13,** 367–371.

Mathews, R. M., & Fawcett, S. B. (1984). Building the capacities of job candidates through behavioral instruction. *Journal of Community Psychology,* **12,** 123–129.

Michael, J. (1980). Flight from behavior analysis. *The Behavior Analyst,* **3,** 1–22.

Morris, E. K. (1982). Some relationships between interbehavioral psychology and radical behaviorism. *Behaviorism,* **10,** 187–216.

Petroski, H. (1985). *To engineer is human: The role of failure in successful design.* New York: Saint Martin's Press.

Pierce, W. D., & Epling, W. F. (1980). What happened to analysis in applied behavior analysis? *The Behavior Analyst,* **3,** 1–9.

Powell, J. (1984). On the misrepresentation of behavioral realities by a widely practiced direct observation procedure: Partial interval (one-zero) sampling. *Behavioral Assessment,* **6,** 209–219.

Reppucci, N. D., & Saunders, J. T. (1974). Social psychology of behavior modification. *American Psychologist,* **29,** 649–660.

Schriner, K. F., & Fawcett, S. B. (in press). Development and validation of a community-concerns report method. *Journal of Community Psychology.*

Seekins, T., Mathews, R. M., Fawcett, S. B., & Jones, M. L. (in press). A market-oriented strategy for applied research in independent living. *Journal of Rehabilitation.*

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

Sidman, M. (1986). Functional analysis of emergent verbal classes. In T. Thompson & M. D. Zeiler (Eds.), *Analysis and integration of behavioral units.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis,* **10,** 349–367.

Ulman, J. D., & Sulzer-Azaroff, B. (1975). Multielement baseline design in educational research. In E. Ramp & G. Semb (Eds.), *Behavior analysis: Areas of research and application.* Englewood Clifs, NJ: Prentice-Hall.

Wahler, R. G., & Fox, J. J. (1982). Response structure in deviant parent-child relationships: Implications for family therapy. In D. J. Bernstein (Ed.), *Response struc-*

ture and organization: The 1981 Nebraska symposium on motivation. Lincoln, NE: University of Nebraska Press.

Wolf, M. M. (1978). Social validity: The case for subjective measurement, or how behavior analysis is finding its heart. *Journal of Applied Behavior Analysis,* **11**, 203–214.

Wolf, M. M., Braukmann, C. J., & Ramp, K. A. (1987). Serious delinquent behavior as part of a significantly handicapping condition: Cures and supportive environments. *Journal of Applied Behavior Analysis,* **20**, 347–359.

Zuriff, G. E. (1985). *Behaviorism: A conceptual reconstruction.* New York: Columbia University Press.

Received May 21, 1987
Revision received July 29, 1987
Final acceptance September 2, 1987
Action Editor, Jon S. Bailey