JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1978, 11, 203-214 NUMBER 2 (SUMMER 1978)

SOCIAL VALIDITY: THE CASE FOR SUBJECTIVE MEASUREMENT or HOW APPLIED BEHAVIOR ANALYSIS IS FINDING ITS HEART[1]

MONTROSE M. WOLF

UNIVERSITY OF KANSAS

I apologize, but I must begin making my case for subjective measurement by recounting to you my own experiences with it over the past few years. Almost a decade ago, when the field of applied behavior analysis was beginning to expand so rapidly, we were faced with the task of putting together the *Journal of Applied Behavior Analysis*. For a period of several months Garth Hopkins, who was our managing editor, presented us with a series of unexpected decisions to make; like: What color should the paper be? And did we need a paper that would hold together for two thousand years or were we willing to live with a shelf-life of only a thousand years? And so on.

Just a couple of days before we were scheduled to go to press with our very first issue, Garth called with one more question. "What is the purpose of the *Journal of Applied Behavior Analysis?*", he asked. He said we needed to put a description of the purpose on the inside front cover, as one finds in other journals. He needed an answer almost immediately.

What was the purpose of our journal? It was a question that was clearly more important than the others I had been asked. So I decided to consult the Gods but, as usual, Don Baer, Don Bushell, Barbara Etzel, Vance Hall, Bill Hopkins, Judy LeBlanc, Keith Miller, Todd Risley, and Jim Sherman were not in their offices. However, I did find Don Baer in the hall. So I asked Don, "What is the purpose of *JABA?*" and Don said in his usual offhand but eloquent way, "It is for the publication of applications of the analysis of behavior to problems of social importance." Well, that sounded so reasonable that it had to be true. So that is what I put in the *Journal* and it went to press.

There was only one small problem; I wasn't sure what "social importance" meant or, worse still, how to measure it. And, as I am sure you can appreciate, the more I thought about this the more concerned I became.

The dictionary only added to my distress. According to my *New Webster's Vest Pocket Dictionary* (1962) importance simply meant "having value" and of course, social meant "pertaining to society". Thus, something of social importance would have to be judged by someone as having value to society.

Unfortunately, that sounded slightly subjective to me. And subjective criteria have not been very respectable in our field. We have considered ourselves a natural science, concerned about the objective measurement of natural events such as arithmetic problems worked correctly, litter picked up, sexual responses occurring, and social skills learned. We have considered ourselves to be like the other natural sciences: like physics, chemistry, and biology, which concern

[1]This manuscript was presented as an invited address to the Division of the Experimental Analysis of Behavior, American Psychological Association, Washington, D.C., September, 1976. Many valuable suggestions regarding this manuscript were made by Don Baer, Curt Braukmann, Steve Fawcett, Dean Fixsen, Bill Hopkins, Frances Horowitz, Kathi Kirigin, Jack Michael, Keith Miller, Todd Risley, Jim Sherman, and Sandra Wolf. Preparation of the manuscript was partially supported by Grants MH20030, MH13644, and MH13881 from the National Institute of Mental Health (Center for Studies of Crime and Delinquency) to the Department of Human Development and the Bureau of Child Research, University of Kansas. Reprints may be obtained from Montrose M. Wolf, Department of Human Development, University of Kansas, Lawrence, Kansas 66045.

themselves with the objective aspects of nature and profitably abandoned the subjective dimensions of natural events sometime in their primordial past.

We have considered ourselves to be distinctly purer and more objective than most of our sister social sciences. We have looked especially askance at our colleagues in sociology, anthopology, psychiatry, and humanistic psychology because they often mix into their sciences difficult-to-digest portions of subjective measurement.

But psychologists have not always been so suspicious of subjective data. For some time, and until the first decades of this century, introspection was the basic method of psychology. As you no doubt remember from your history of psychology course, introspection is defined as the observation or examination of one's own mental, emotional, or feeling states. The subjects' verbal descriptions about sensations, private events, and feelings such as pleasantness and unpleasantness had been taken to be the primary subject matter of psychology (Boring, 1950). As a reaction against introspection in psychology and in science generally, there arose positivism from Bridgeman in physics and from Comte, Mach, and Feigl in philosophy. To quote Edwin Boring (1950) about its impact:

> "The movement was positivistic. It was an attempt to get back to basic data and thus to increase agreement and diminish the misunderstandings that came about from unsuspected differences in meaning. Experience [introspection] had proved unsuccessful as the scientific ultimate." (Boring, 1950)

John Watson began page one of his book *Behaviorism* in the following manner:

> "Two opposed points of view are still dominant in American psychological thinking —introspective or subjective psychology, and behaviorism or objective psychology. Until the advent of behaviorism in 1912, introspective psychology completely dominated American university psychological life." (Watson, 1930).

B. F. Skinner, in *Science and Human Behavior* (1953), also argued forcefully against subjective measures of private events. He began by pointing out the implications of the discriminated operant model of language. He described how a community can reinforce and thus develop reliable verbal reporting of public events because both the community and the individual have access to these events. On the other hand, he pointed out that since the community cannot have access to private events, the use of psychology of introspective or subjective data leads to serious questions about reliability. Skinner continued,

> "The layman also finds the lack of a reliable subjective vocabulary inconvenient. Everyone mistrusts verbal responses which describe private events. Variables are often operating which tend to weaken the stimulus control of such descriptions, and the reinforcing community is usually powerless to prevent the resulting distortion. The individual who excuses himself from an unpleasant task by pleading a headache cannot be successfully challenged, even though the existence of the private event is doubtful."

While defining a functional analysis for us, Skinner (1953) urged us to concentrate on the objective behavioral data in our science as in the following quotation:

> "The objection to inner states is not that they do not exist, but that they are not relevant in a functional analysis. . . . In dealing with the directly observable data we need not refer to . . . the inner state. . . ."

Having been well trained in these traditions, we all agreed that in our journal, everything would be measured in objective ways. We would avoid subjective measurement—that would be a first priority. Some of the members of the

JABA Board of Editors even wanted to restrict us to using only mechanically recordable behavior in our applied research. They wanted a microswitch under every schoolroom chair and under every bed. They were even suspicious of observer measurement systems that contained reliability checks. Yet I, in a moment of haste, had committed our journal to a goal, to an ultimate criterion, to a reason for being, that was clearly and simply subjective and that we had no good way of measuring.

You can imagine what I expected. I prepared for an onslaught of abuse, invective, and ridicule from our editors and our reading audience. "Social importance? Bah! Humbug!", I thought they would say. To my surprise and relief, what happened was that people seemed pretty much to accept it. Many even seemed to know what it was. For example, *JABA* editors often referred to it in their reviews and used it as a basis for recommending or not recommending manuscripts for publication. The editors most frequently reported that the particular manuscripts that they had been asked to review didn't have very much of it. On the other hand, they reported that a few manuscripts had a moderate amount of it. And an occasional one or two had a lot of it. This made me feel somewhat better. Although I wasn't sure what it was or how to measure it objectively, it was clear that many of my colleagues had no trouble at all in recognizing it.

I was also fearful of criticism from our reading audience. And we did receive occasional complaints about social importance. But primarily they wanted to know why the research that appeared in *JABA* was not *more* socially important. That criticism was easy for me to live with. I just blamed our authors. If the readers had taken me to task for using a fuzzy subjective criterion like "social importance", then I would have had no excuse.

But the issue of subjective measurement continued to make my life complicated. One of the functions of a chief editor is to uphold the standards of the journal. And almost everyone in the field strongly suggests that these be maintained rigorously. Except, of course, in the special case of everyone's own manuscripts which, because of their unusual significance, merit special consideration. In any event, among the standards that I was entrusted to uphold was that of requiring objective, reliable data. Thus, you can appreciate the concern I began to feel when some of our most esteemed colleagues began submitting articles to *JABA* that included undisguised, blatantly subjective data.

One of the first came from, of all people, Bob Jones and Nate Azrin (1969). They had been conducting an exquisite series of experiments on the effects of rhythm and stimulus duration on stuttering behavior. They had shown, very nicely, that they could almost completely eliminate stuttering by having the stutterers synchronize their speech with a simple, regular beat. They had also developed a portable practical piece of apparatus that would present the beat tactually, and privately, thus avoiding embarrassment to the wearer. Their results indicated that they were on the verge of an important solution to stuttering. There had been one problem, however. The speech, although almost stutter-free, was complained about by listeners as sounding *artificial.* [The next sentence is to be read with a monotone with a distinct beat.] Apparently, they did not stutter, but they did not talk very naturally, either.

To deal with this problem, Jones and Azrin systematically explored various beat durations. Then,—and this was the difficult part—they asked judges to rate the "naturalness" of the speech at various beat durations. The judges reported that the speech sounded most natural to them at between two and three seconds of beat duration.

I wanted to phone Jones and Azrin and say, "Hey you guys, do you realize what you are doing to me and the journal? Do you realize what kind of precedent you will be setting with your 'naturalness'? Why, the people in our field who are not as sophisticated as you and me and who are easily influenced will begin to think that it

is possible to measure how people feel about all kinds of subjective things. I know that 'naturalness' sounds innocent enough, but think about it a moment. If you publish a measure of 'naturalness' today, why tomorrow we will begin seeing manuscripts about happiness, creativity, affection, trust, beauty, concern, satisfaction, fairness, joy, love, freedom, and dignity. Who knows where it will end? Think for just a moment. What is that going to do to us and to the field of applied behavior analysis?"

But I was sure that they would have just said that they would agree that it was going to complicate our science a bit. But if those things described by subjective labels were the things that were most important to people, then those were the things, even though they might be complex, that we should become more concerned with. After all, as an applied science of human behavior, we supposedly were dedicated to helping people become better able to achieve their reinforcers.

Well, it didn't stop with Jones and Azrin. At about the same time I received a lovely manuscript from Jim McMichael and Jeff Corey (1969) in which they reported the exciting finding that college students in a Keller-type PSI (Personalized System of Instruction) course did better on the exam than the students in a traditional lecture course. This was, of course, a very important finding, as it replicated and substantiated Keller's research. The only problem was that they also asked the students in each course how much they liked their course. The students in the PSI course rated their course a great deal higher than the students in the traditional lecture sections.

"Well," I thought to myself, "What in the world am I going to do with this one? They are asking the participants in a behavioral treatment program how much they like it. Why, of course they should like it. After all, we are doing it to them for their own good aren't we? And even if they say they don't like it, we know what is best for them. Clearly, if the procedure is effective, its just not important whether anyone says they like it or not. Besides, look at the precedent that it will set. Before long, those who don't appreciate the extreme risks of subjective data will start asking for feedback from the participants in their treatment programs. Who knows where that will end?"

But I felt sure that McMichael and Corey would just say that feedback from participants is not a trivial issue: that if the participants don't like the treatment then they may avoid it, or run away, or complain loudly. And thus, society will be less likely to use our technology, no matter how potentially effective and efficient it might be.

At the same time that I was having to wrestle with the problems of subjective measurement in *JABA,* my colleagues and I in the Achievement Place Research Project were having some problems with unsolicited subjective feedback on similar issues. Colleagues, editors, and community members were asking us about the behavioral goals that we had chosen for training the teaching-parents and the youths participating in the community-based, family-style, behavioral treatment program at Achievement Place. They would ask us: "How do you know what skills to teach? You talk about appropriate skills this and appropriate skills that. How do you know that these are really appropriate?" We, of course, tried to explain that we were psychologists and thus the most qualified judges of what was best for people. Somehow, they didn't seem convinced by that logic.

In addition, the first time we tried to replicate the Achievement Place program in another community, that community gave us feedback in a most drastic manner. Before we really knew that they had complaints about our program they had "fired" us. Finally, there were those who were challenging the importance of some of the results of the training that we were reporting. "Yes," they would say, "there are changes in the behavior, but how do we know that they are really important changes?"

The message we seemed to be getting was that "social importance" was a subjective value

judgement that only society was qualified to make. If our objective was, as described in *JABA,* to do something of social importance, then we needed to develop better systems and measures for asking society whether we were accomplishing this objective. The suggestion seemed to be that society would need to validate our work on at least three levels:

1. The social significance of the *goals.* Are the specific behavioral goals really what society wants?
2. The social appropriateness of the *procedures.* Do the ends justify the means? That is, do the participants, caregivers and other consumers consider the treatment procedures acceptable?
3. The social importance of the *effects.* Are consumers satisfied with the results? *All* the results, including any unpredicted ones?

We have come to refer to these as judgements of *social validity.* It seems to us that by giving the same status to social validity that we now give to objective measurement and its reliability we will bring the consumer, that is society, into our science, soften our image, and make more sure our pursuit of social relevance.

An example from our own experience in the Achievement Place Research Project is that we were told by many communities that one of the most important characteristics of teaching-parents that they wanted was "warmth". When quizzed about "warmth", the community members indicated that they wanted teaching-parents who "know how to relate to youths". For some time, our response to this request was to disagree with them. We argued, "What you really need is someone who knows how to give and take away points at the right time." But the results of our research (Braukmann, Kirigin, and Wolf, 1976) are tending to support the community's commonsense wisdom about the importance of teaching-parents being able to "relate to youths".

Thus, in order to be responsive to our communities and to our data, one of our challenges became to try to determine the behaviors that teaching-parents need in order to "relate to their youths". "What do some people have that makes kids like them? And how were we going to find out?", we asked ourselves over and over. "Relating" appeared to be such a complex behavioral puzzle of subtle social behaviors that we were not sure how to begin our behavioral analysis. We did have the Jones and Azrin example for measuring "naturalness", and we came upon another method from, of all places, the Rogerian counselling psychologists.

Haase and Tepper published an article in the *Journal of Counseling Psychology* in 1972 that was a great deal of help to us. Like so many Rogerians, Haase and Tepper were interested in "empathy". They wanted to see if they could find out what nonverbal behaviors of the counsellor were involved in empathy in order to be better able to teach and evaluate counsellors in training. They set up simulated counselling situations that contained various nonverbal components, such as level of eye contact, trunk lean (forward or backward), body orientation (toward the client or rotated away from the client), distance from the client and various levels of "empathic" verbal messages". Videotaped excerpts were then presented to experienced counsellors, who rated the amount of overall empathy presented in each excerpt. It was found that eye contact, trunk lean, distance, and verbal content were all related to the judgements of empathy. One result that really seemed to surprise the authors was that the nonverbal behaviors accounted for more than twice as much of the judgements of empathy than did the verbal behaviors. A counsellor who was saying something only moderately empathic was judged to be highly empathic if he or she were also engaging in eye contact, forward trunk lean, and were positioned close to the client.

Well, it occurred to us that this model could be used to analyze the *meaning* of all kinds of complex and subjective verbal labels. It also

looked like a way to find out what some of the behaviors were that made some teaching-parents better than others in being able to "relate to youths". Alan Willner, with Curt Braukmann, Kathi Kirigin, Dean Fixsen, Lonnie Phillips, and I (Willner *et al.,* 1977) began to attempt to identify the interaction behaviors of teaching-parents in Achievement Place style group homes the youths liked and didn't like. Alan Willner had several youths look at videotaped examples of a variety of teaching-parent/youth interactions and to list the things that they liked and the things that they disliked. These comments were put into categories and then rated by the youths on an A, B, C, D, and F basis. The youths gave A's to the following teaching-parent behaviors: a calm, pleasant voice tone, offers to help, joking, fairness, explanations, concern, enthusiasm, politeness, and getting to the point. F's were given to the following teaching-parent behaviors: throwing objects, accusing, blaming statements, shouting, no opportunity provided to speak, insulting remarks, unfair point exchanges, and profanity. Willner then took some of the highest rated social behaviors, taught them to teaching-parent trainees, and found that youths rated these trainees much higher after the trainees received instruction in the youth-preferred behaviors.[2]

One important sidelight of Alan Willner's study was that he was not able to predict the behaviors of the teaching-parents that were going to be most liked by the youths. As a matter of fact, some of the behaviors that he thought would be most important to the youths were never mentioned by them. He still wasn't convinced. After all, maybe the youths just couldn't verbalize these subtle behaviors—which of course was a real possibility. In this case, however, he cross-validated the original behaviors by giving the youths more structured interviews, in which he included more detailed descriptions of the behaviors that he thought should also be important to them. The youths still rated those behaviors as much less important than the ones that they had earlier pointed out as important. This same outcome was found with youths who were not involved in the first set of interviews. It has become clear to us that we cannot predict very well what many subjective labels of complex behavioral phenomena are going to mean to our judges. Nevertheless, while the task of unravelling those social behaviors that are involved in knowing how to "relate to youths" is incomplete, Alan Willner has taken us closer to that goal.

Another example of the use of the social validation method to examine the social validity of behavioral goals is a study by Neil Minkin, with Curt Braukmann, Bonnie Minkin, Gary Timbers, Barbara Timbers, Dean Fixsen, Lonnie Phillips—and me (Minkin *et al.,* 1976). Neil Minkin wanted to determine what conversational skills of adolescent girls were relevant. He took videotapes of adolescent girls in conversations with adults and of university girls in conversations with adults. Judges from the community were then asked to rate the effectiveness of each of these girls as conversationalists. As might be expected, the community people judged the university girls to be more effective and ranked them higher. Minkin and others reviewed the videotapes of all the university and junior high-school girls several times, and determined that a composite score of three kinds of behavior correlated at the 0.84 level with

[2]Jack Michael (personal communication, 1976) has pointed out that some behaviors, identified as preferred by this method, may have acquired their reinforcing value by their usually being members of chains of behaviors. An example might be *offers to help.* It is possible that if *offers to help* were not often followed by *providing help,* the offers themselves would lose their reinforcing value. Similarly, behaviors described as showing *concern* may have the same relationship to a more complex chain of behaviors. Thus, there appears to be an important and not, as yet, well understood "sincerity" dimension that should be brought to the attention of anyone who may want to apply these findings. On the other hand, some of the behaviors identified as preferred may not be dependent on later events for their reinforcing value. Examples might be *joking* and *explanations.*

the ratings given by the community representatives. (The three behaviors were: time spent talking, conversational questions, such as "What are you taking in school?", and positive feedback behaviors such as "Uh huh", "Yeah", and "Great!") In this manner it was possible to isolate many of the behaviors that the community representatives clearly were responding to when they rated overall quality of a conversation.

Another example of the social validation of behavioral goals, conducted by the Achievement Place group, was carried out by Jack Werner, with Neil Minkin, Bonnie Minkin, Dean Fixsen, Lonnie Phillips—and me (Wérner *et al.*, 1975). Police exercise a great deal of discretion in handling juvenile offenders. Less than one-fourth of those youths who come into contact with police officers and who could be taken into custody actually are taken into custody. According to Piliavin and Briar (1964), the violation *per se* is usually less influential in determining the choice of disposition than is the demeanor of the youth. It is often estimated that the social behaviors of the youths account for approximately 50% of all decisions regarding prejudicial handling of youths. Jack Werner wanted to identify some of the important behavioral components of youth-police interactions so that he could teach these to youths. Through informal interviews and then formal questionnaires, Werner and his colleagues identified several apparently important behaviors, including expression of cooperation, body orientation so that the youth was facing the officer, and politeness. Werner found that these behaviors could be reliably measured, thus partially solving the behavioral puzzle of what objectively measurable youth behaviors may influence police officers' decisions about custody.

So, rather than deciding by oneself the validity of the behavioral objectives of a treatment program, we can approach the specific consumer or representatives of the relevant community, and through interviews or ratings determine much more precisely what the socially significant problems are. And, based on the example of Jones and Azrin (1969) and the work of Haase and Tepper (1972), we find that we can establish the social importance or validity of complex classes of behavior that have subjective labels. By supplementing our traditional objective measures, we can determine the relationship between the objectively measured behaviors and the subjective labels. This procedure opens opportunities to explore all of the important goals that are described by subjective labels.

To summarize the method for determining goal behaviors, I quote from Minkin and his colleagues (1976):

> "For example, 'affection' might be considered a complex social behavior. If the goal of a behavior analyst were to teach a parent to be more affectionate towards his or her child, it would be necessary to specify the important component behaviors of affection. Some of the components might include touching, smiling, and hugging. To validate the social importance of these behaviors, four steps might be used. First, gathering sample parent-child interactions. Second, developing reliable definitions and recording specific behaviors. Third, employing relevant judges, that is, other parents or children, to rate the sample interactions and evaluate each parent as to the amount of affection shown to the child within the interaction. The evaluation instrument might be a bi-polar rating scale with the poles labelled as to the amount of affection shown. Step four would involve correlating the ratings of the judges with a composite score of the objectively measured behaviors of the parents. The subsequent correlation coefficient would indicate the level of relationship of the specified objectivity measured components of affection to the common English 'meaning' of affection as rated by the judges. Some of the important behavioral components of creativity, conversation, and affection, as well as other complex classes of social

behaviors, could probably be identified through the use of these social validation procedures."

It is clear that a number of the most important concepts of our culture are subjective, perhaps even the most important. Martin Luther, as the story goes, was severely criticized for setting Potestant hymns to the popular melodies of songs and dances of the time. He replied, "Why should we let the devil have all the best tunes?" Well, why should we let the others have all of the best human goals and social problems?

A second kind of social validity that has impressed its importance on us is the social appropriateness (in terms of ethics, cost, and practicality) of the treatment procedures that we use. Again, behavior analysts are beginning to ask clients and care-givers systematically about the acceptability of their procedures. Foxx and Azrin (1972) found restitution procedures more acceptable to care-givers than timeout or shock punishment. These authors have also reported over-correction to be a re-education procedure that is acceptable to care-givers of the retarded.

Janet Porterfield, Emily Herbert-Jackson, and Todd Risley (1976) recently determined that "contingent observation" (that is, having to stop playing and just watch your playmates for several seconds) was not only an effective procedure for reducing the disruptive behavior of young children in a day-care setting, it was also found to be acceptable to the care-givers and to the parents of the children.

Our own data show that ratings by the youths in Achievement Place style homes of the fairness of the program and the concern of the teaching-parents correlate very highly with the number of offenses that the youths commit while they are in treatment (Braukmann, Kirigin, and Wolf, 1976). It may be that not only is it important to determine the acceptability of treatment procedures to participants for ethical reasons, it may also be that the acceptability of the program is related to effectiveness, as well as to the likelihood that the program will be adopted and supported by others.

The third dimension of social validity is the social importance of the effects of behavioral treatment. Are consumers satisfied with the results, all of the results, including those that were unplanned? Behavioral treatment programs are designed to help someone with a problem. Whether or not the program is helpful can be evaluated only by the consumer. Behavior analysts may give their opinions, and these opinions may even be supported with empirical objective behavioral data, but it is the participants and other consumers who want to make the final decision about whether a program helped solve their problems. Many behavior analysts are beginning to validate their objective data with systematic subjective measures of consumer satisfaction.

For example, Ron Kent and Dan O'Leary (1976) found the ratings by teachers and parents of child behavior also improved when their objective data showed increases in appropriate school behavior. Karen Maloney and Bill Hopkins (1973) determined that when they modified the sentence structure of stories written by elementary school children, judges' ratings of creativity also increased. This is to be contrasted with the findings of Tom Brigham, Paul Graubard, and Aileen Stans (1972), who were also attempting to improve quality of composition of school children, and found that some contingencies that increased objective dimensions had little effect on subjective ratings of quality, while other contingencies produced increases in both objective measures and subjective ratings of story quality. Steve Fawcett and Keith Miller (1975) demonstrated that an instructional package designed to enhance public-speaking behavior was effective in producing increases in both the objectively measured public-speaking behaviors and in the audience's ratings of the quality of the performance of the trainees.

We have described the Achievement Place research of Willner, Minkin, and Werner and their colleagues, where judges were used to de-

termine socially valid dimensions of teaching-parent/youth interaction behavior, quality of conversation components, and significant elements in youth-police interaction. In each of those studies, the outcomes were also socially validated. That is, relevant judges were also used to assess the social importance of the changes in the objectively measured behaviors. And it was found that youths rated the quality of the teaching-parents higher, members of the community rated the quality of the youths' conversations higher, and police officers rated the quality of the demeanor of the youths higher as the objectively measured behaviors increased in each case.

At the treatment program level, Curt Braukmann with Dean Fixsen, Kathi Kirigin, Elaine Phillips, Lonnie Phillips—and I (1975) described how feedback from consumers can be used to provide ongoing quality control of the dissemination of the Achievement Place treatment model. The consumers of the program, that is the youths in the program, their parents, and community members and agencies, evaluate the teaching-parents by rating their effectiveness, concern, *etc.* throughout the year of training and certification, and each year thereafter. It has not been possible to demonstrate experimentally the effectiveness of this feedback system by using it with some programs and not with others because of ethical considerations. But there is one important bit of data. Since this feedback was put into effect, the Achievement Place program has not been summarily "fired" from a community, as in that first attempt at replication. Also, these consumer satisfaction ratings are often highly correlated with objective measures of effectiveness (Braukmann *et al.,* 1976).

Concern for the social validity of *objective* measures seems to be an issue in other social sciences as well. At the American Psychological Association meeting, Angus Campbell (1976) raised this issue about economics:

> "None of us doubts that economic data have admirable qualities: the question is, How well do they represent the quality of national life? How valid are they as measures of the goodness of life in this country? The history of the last 25 years is not reassuring. During this period this country has experienced an unprecedented rise in national affluence, with a spectacular increase in average family income and an associated decline in the number of families below the poverty line. During the same period we have seen a phenomenal rise in the incidence of crime, an epidemic of various forms of public violence, a greatly increased use of drugs with associated drug abuse, a continuing increase in the number of fragmented families, a sharp drop in public confidence in elected officials, and what appears to be a substantial rise in social and political alienation. [I] . . . find it hard to believe that the quality of American life has been greatly enhanced during this period."

E. F. Schumaker, in his book *Small is Beautiful: Economics as if People Mattered* (1973), raised this same issue. He urged economists to consider what he terms the *"primacy of qualitative distinctions"*, rather than being so concerned with objective data like the gross national product.

Recently, the Swedish medical sociologists Levi and Anderson (1975) suggested that objective measures that habitually have been used by the United Nations to assess the quality of life be supplemented by subjective measures. They proposed that the traditional objective measures of quality of life, such as education, employment, economy, housing, nutrition, *etc.* be given equal emphasis with subjective criteria such as "happiness, satisfaction, and gratification". Thus, applied behavior analysts are not the only applied social scientists who are being asked to validate their measures by checking with society.

Well, if social validity is such a good thing, why haven't we been doing more of it all along.

Of course, the answer is that subjective data are risky data. Subjective data may not have any relationship to actual events. A program that is described by its consumers as well-liked or effective may not necessarily be either pleasant or effective. Thus, there is the danger that subjective data will seriously mislead us.

For example, Berleman, Seaberg, and Steinburn (1972) conducted a delinquency prevention experiment with carefully matched experimental and control groups, using intensive one- to two-year treatment by social workers as the intervention procedure. The evaluation of the effectiveness during the treatment period and during the eight months following treatment indicated "no positive impact" on disruptive behavior in school, police contacts, or rate of institutionalization. The untreated control group performed as well or better than the experimental group. Yet, when asked about their experience in treatment, the youths ". . . believed that their school acting-out had decreased. When asked if they would participate in a similar service again, 89 percent of the parents responded positively, as did 94 percent of the boys".

Behavioral researchers have reported many examples of a lack of correspondence between client-reported data and observer-obtained data. Patterson (*personal communication,* 1974) for example, described discrepancies between parental reports of improvements in the child's behavior, while objective data obtained by observers did not support these claims. Conrad and Wincze (1976) reported that clients undergoing orgasmic reconditioning verbally reported favorable results that were not substantiated by the objective data.

Why do these discrepancies exist? One possibility is that the contingencies of the situation create distortion. Verbal behavior, clearly, is a manipulable behavior. And we must be suspicious of it because we know that we will not always understand the contingencies operating on it. When we are asking for a verbal description of a private event, such as satisfaction with our treatment program, we must be very cautious because we have no adequate way of checking the reliability of the verbal report in an independent way. And as Skinner pointed out, verbal descriptions of private events are open to "fictional distortion" (1959).

For example, in order to influence consumer evaluations, it is conceivable that some of those being evaluated might politic their consumers for better ratings. Similarly, it is conceivable that some of those consumers giving ratings might fear that they will not remain anonymous and be afraid that those they are rating might retaliate in some manner. One can conceive of many such possibilities, but let us remember that the reliability of objective measurement systems can also be manipulated, as the excellent series of studies by O'Leary and Kent and their colleagues (O'Leary, Kent, and Kanowitz, 1975; Kent, O'Leary, Diament, and Dietz, 1972) have demonstrated. From these studies, it seems clear that the scoring behavior of observers can be affected by a variety of variables, such as experimenter feedback. We must take these into consideration whenever we design a measurement system that involves observers. Thus, we know that the reliability of objective measurement procedures can be influenced by a number of known and probably unknown variables, but we continue to use these systems because they are the only way to obtain some very important data, they often work, and we feel some confidence that we are gaining a better understanding of the conditions that may distort them.

Similarly, we know that social validity measures can be manipulated and abused, but we cannot allow this to lead us to neglect them. Rather, we must establish that set of conditions under which people can be assumed to be the best evaluators of their own treatment needs, procedural preferences, and posttreatment satisfaction. True, we know little about the proper set of conditions, but we must attempt them anyway. We can expect that they will involve education about options, lack of coercion, an-

onymity, and so on. We can study the effects of these conditions on subjective data, as O'Leary and Kent and their colleagues have studied their effects on objective observer-dependent measurement systems. And then we will be better able to control for them.

A second possible explanation for subjective-objective discrepancies is that the consumer is responding to changes in some behavior or condition that we are not recording with our particular objective measures. For example, the parent may say that a child has "improved", while our behavioral measure of rate of tantrums does not show a decrease. The discrepancy may be because the child has stopped cursing, which was important to the parent, but not measured by us, perhaps because it does not bother us. If this lack of appropriate measurement is one of the factors in subjective-objective discrepancies, then we must become better at setting up our measurement systems.

A third possibility, and the most serious, is that subjective measurement is impossible because humans cannot judge and report their own situation accurately enough. It may be that they don't know when they are better or worse off. It may be that to expect a human ever to be able to report accurately when something feels good or feels bad is just more than we can hope for from our confused species. But this conclusion is unacceptable if our goal is to design a responsive consumer-oriented applied social science. As Levi and Anderson (1975) argued in making their case for adding subjective measures to objective quality-of-life indicators:

> "We believe that each individual can be assumed to be the best judge of his own situation and state of well-being. The alternative is some type of 'big brother' who makes the evaluation for groups and nations. World history provides many examples of such 'expert' or 'elitist' opinions being at variance with what was expected by the man in the street."

Therefore, we may have to try to develop better ways of teaching people to observe their behavior and their conditions and to make more accurate decisions about their improvement. The opinion poll people often seem to be able to make excellent predictions about voting behavior based on verbal report. Surely we can do as well.

Undoubtedly, there will be further important studies that point out to us the shortcomings of certain social validity measures, just as has been done for observer-dependent objective measures. But we can't despair. After all, measurement has been our thing. In our field, we have developed so many ingenious measurement systems. There is no doubt that we could measure the disruptive classroom behavior of a school of fish, if need be. Surely, we will be able to develop measurement systems that will tell us better whether or not our clients are happy with our efforts and our effects.

Earlier in our history, Watson and Skinner argued forcefully against subjective measurement because they were concerned about the inappropriate causal roles that hypothetical internal variables, subjectively reported, were playing in social science. As a result, many of us concluded that all subjective measurement was inappropriate. A new consensus seems to be developing. It seems that if we aspire to social importance, then we must develop systems that allow our consumers to provide us feedback about how our applications relate to their values, to their reinforcers. This is not a rejection of our heritage. Our use of subjective measures does not relate to internal causal variables. Instead, it is an attempt to assess the dimensions of complex reinforcers in socially acceptable and practical ways. It is an evolutionary event that is occurring as a function of the contingencies of the applied research environment; contingencies that our founders would probably say they appreciate, if we had the nerve to ask them for such subjective feedback on our behavior.

REFERENCES

Berleman, W. C., Seaberg, J. R., and Steinburn, T. W. The delinquency prevention experiment of the

Seattle Atlantic Street Center: A final evaluation. *Social Science Review,* 1972, Sept., 323-346.

Boring, E. G. *A history of experimental psychology.* New York: Appleton-Century-Crofts, 1950.

Braukmann, C. J., Fixsen, D. L., Kirigin, K. A., Phillips, E. A. Phillips, E. L., and Wolf, M. M. Achievement Place: The training and certification of teaching-parents. In W. S. Wood (Ed), *Issues in evaluating behavior modification.* Champaign, Illinois: Research Press, 1975. Pp. 131-152.

Braukmann, C. J., Kirigin, K. A., and Wolf, M. M. *Achievement Place: The researchers' perspective.* Paper presented at the meeting of the American Psychological Association, Washington, D.C., September, 1976.

Brigham, T. A., Graubard, P. S., and Stans, A. Analysis of the effects of sequential reinforcement contingencies on aspects of composition. *Journal of Applied Behavior Analysis,* 1972, **5,** 421-430.

Campbell, Angus. Subjective measures of well-being. *American Psychologist,* 1976, **31,** 117-124.

Conrad, S. R. and Wincze, J. P. Orgasmic reconditioning. A controlled study of its effects upon the sexual arousal and behavior of adult male homosexuals. *Behavior Therapy,* 1976, **7,** 155-166.

Fawcett, S. B. and Miller, L. K. Training public-speaking behavior: an experimental analysis and social validation. *Journal of Applied Behavior Analysis,* 1975, **8,** 125-136.

Foxx, R. M. and Azrin, N. A. Restitution: A method of eliminating aggressive-disruptive behavior of retarded and brain damaged patients. *Behaviour Research and Therapy,* 1972, **10,** 15-27.

Hasse, R. F. and Tepper, D. T. Nonverbal components of empathetic communication. *Journal of Counseling Psychology,* 1972, **19,** 417-424.

Jones, R. J. and Azrin, N. A. Behavioral engineering: stuttering as a function of stimulus duration during speech synchronization. *Journal of Applied Behavior Analysis,* 1969, **2,** 223-230.

Kent, R. N. and O'Leary, D. K. A controlled evaluation of behavior modification with conduct problem children. *Journal of Consulting and Clinical Psychology,* 1976, **44,** 586-596.

Kent, R. N., O'Leary, K. D., Diament, C., and Dietz, A. Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology,* 1972, **42,** 774-780.

Levi, L. and Anderson, L. *Psychosocial stress: Population, environment, and the quality of life.* Holliswood, N.Y.: Spectrum Press, 1975.

Maloney, K. B. and Hopkins, B. L. The modification of sentence structure and its relationship to subjective judgements of creativity in writing. *Journal of Applied Behavior Analysis,* 1973, **6,** 425-434.

McMichael, J. S. and Corey, J. R. Contingency management in an introductory psychology course produces better learning. *Journal of Applied Behavior Analysis,* 1969, **2,** 79-84.

Minkin, N., Braukmann, C. J., Minkin, B. L., Timbers, G. D., Timbers, B. J., Fixsen, D. L., Phillips, E. L., and Wolf, M. M. The social validation and training of conversation skills. *Journal of Applied Behavior Analysis,* 1976, **9,** 127-140.

New Webster's Vest Pocket Dictionary. Ottenheimer Publishers, Inc., 1962.

O'Leary, K. D., Kent, R. N., and Kanowitz, J. Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behavior Analysis,* 1975, **8,** 43-51.

Piliavin, I. and Briar, S. Police encounters with juveniles. *American Journal of Sociology,* 1964, **70,** 206-214.

Porterfield, J. K., Herbert-Jackson, E., and Risley, T. R. Contingent observation: an effective and acceptable procedure for reducing disruptive behavior of young children in a group setting. *Journal of Applied Behavior Analysis,* 1976, **9,** 55-64.

Schumaker, E. F. *Small is beautiful: economics as if people mattered.* New York: Harper & Row, 1973.

Skinner, B. F. *Science and human behavior.* New York: Macmillan Co., 1953.

Skinner, B. F. *Cumulative record.* New York: Appleton-Century-Crofts, Inc., 1959.

Watson, John B. *Behaviorism.* Chicago: The University of Chicago Press, 1930.

Werner, J. S., Minkin, N., Minkin, B. L., Fixsen, D. L., Phillips, E. L., and Wolf, M. M. Intervention package: An analysis to prepare juvenile delinquents for encounters with police officers. *Criminal Justice and Behavior,* 1975, **2,** 55-83.

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

Received 15 October 1976.
(Final Acceptance 12 August 1977.)