METHODOLOGICAL, MEASUREMENT AND SOCIAL ACTION CONSIDERATIONS RELATED TO THE ASSESSMENT OF LARGE-SCALE DEMONSTRATION PROGRAMS

Clarence C. Sherwood, Action for Boston Community Development, Inc.

The roles which the social scientist can and to some extent is playing in influencing social action are, in my opinion, increasing in both variety and in importance. To a considerable extent, the effective performance of those roles will depend upon the availability of valid, applicable knowledge concerning which social action programs, or interventions, work and which do not. I am personally convinced that the most rational and in the long run the most effective way of accumulating such knowledge will be by conducting action programs as controlled experiments in the community.

I am well aware that the day when such findings will play a dominant part in the decisions concerning the modification, expansion or discontinuance of action programs is probably not in the immediate future. But, I also believe that there is a rapidly developing realization within broad sectors of our population that this kind of knowledge is needed and that the basic questions concerning what works and what doesn't work will be raised with increasing frequency. It is generally recognized, however, that at the present time the social sciences do not have a vast reservoir of findings which are clearly applicable for making decisions about how to attack broad scale social problems. What may not be as clearly or easily recognized is that the social sciences have not had a great deal of experience with large scale social-action programs nor have they thought through how to go about conducting such programs in a way that reasonably hard findings concerning their efficacy can be extracted. We are in a relatively infant stage not only with respect to knowledge of the effectiveness of action programs but also with respect to solutions of the conceptual, theoretical and methodological problems pertaining to the acquisition of such knowledge.

The importance and seriousness of this state of affairs has been augmented many fold by the advent of the multi-billion dollar War on Poverty. Although it is not clear at this moment as to how much emphasis will be placed on the evaluation of anti-poverty programs, there are rumblings that some of the resources will be used to try to ascertain the effects of the attack, or at least of parts of it. I think that social scientists and professionals in related fields should press for a maximum effort in this direction and then organize their own resources toward a major contribution in developing methodologies to fulfill the promises and expectations involved. If some serious effort along these lines is not made, literally billions of dollars will have been spent and we will still not know what works or how to go about finding out about it.

However, no real support for the pursuit of this kind of knowledge is going to be forthcoming until the case for it is presented in practical, useful terms. My recommendation is to point out that a major use of findings from social-action experiments is to provide a basis for a more efficient allocation of financial and human resources to the solution of social problems.

It is this notion of the efficient allocation of resources that I believe is the key to the whole problem of planning and choosing among social-action programs. If this is true, we need a basis for making judgments about efficient allocation, and for this purpose I suggest that we examine very carefully our whole concept of social action or social service programs. Traditionally, service has been viewed--and in a vague way measured--in terms of that which is offered--counseling, guidance, therapy, advice, and the like. Good service is therefore that which is offered in a professional manner by a qualified person who in turn is supervised by a qualified supervisor. My contention is that service must be viewed in terms of impact rather than process. Its success must be viewed in terms of outcome rather than in terms of the quality of the procedures used. The implications of this shift in view are, in my opinion, considerable:

(1) It forces those responsible for program design to clearly specify their objectives, to define what it is they are trying to achieve, what specific changes they are trying to effect At the very least, it requires them to co-operate in efforts to operationalize what they have in mind;

(2) It shifts the emphasis from "procedure as an end" to "procedure as a means." Program personnel must then consider the relationship between the procedures it recommends and the defined outcomes that have been chosen;

(3) It leads to a reconsideration of the whole notion of cost of service. Currently, we are in the grip of the proponents of the "per capita cost of service" point of view. Per capita cost of service is defined as the quotient obtained by dividing the total dollar cost of a program by the total number of individuals exposed to it. If, for example, a particular youth employment program involves 1,000 youths and costs a half million dollars a year, traditional calculations would say that the per capita cost of service is $500. But what if, as well might be the case, only 50 more of the 1,000 youths ended up working steadily (if that is the objective of the program) than would have been the case without the intervention. Calculated on the basis of an impact definition, the per capita unit cost would be $10,000--a vastly different amount. The per capita cost view has resulted in a distinct preference for those programs which "serve" the largest numbers for the least amount of money. If programs were to compete on the basis of how much it costs to achieve one unit (however that may be defined) of desired outcome, our ultimate selection of programs would be, I believe, very different;

(4) And, finally this view forces the inclusion of solid, empirical research into the over-all planning and program operation, because the decisions as to the optimum allocation of the resources available can, within this view, only be made on the basis of empirical evidence.

The challenge is then, at least to me, very clear. Can social experiments be conducted in the community in such a way that the findings resulting therefrom deserve the attention of the policy and decision makers in the community? And what are the current and long-run obstacles to the implementation of action research that prevent it from meeting standards of scientific acceptability?

About two years ago I made the plunge and took on the job of attempting to develop an evaluation program for a large scale community action program, called ABCD (Action for Boston Community Development), in which we are attempting to build research into action program designs. To my knowledge, no similar large scale evaluation research effort had been made before, although variations on the same theme were being developed in New York City, Cleveland, Los Angeles, New Haven, Oakland, Philadelphia, St. Louis, Chicago, Detroit, and a number of other cities and communities. The current state of my thinking is largely the product of my participation in this rather ambitious effort, and therefore, in the analysis which follows I will draw heavily upon my experience for illustrative materials as well as for insights and possible lessons to be learned.

Therefore, before discussing the problems of evaluating large scale social action experiments, let me tell you a little bit about ABCD and the way in which we have tackled this task.

ABCD was envisioned by its founders as the human side of urban renewal. Its first staff people were community organizers and one of their major tasks was to aid in explaining renewal projects to the residents of renewal areas in an effort to enlist their support.

A Ford Foundation grant and a grant from the President's Committee on Juvenile Delinquency and Youth Crime gave ABCD an impetus in another direction. Since then, and until recently at least, our focus has been on developing demonstration programs to be carried out by public and private agencies in the three areas of Boston with the lowest income, highest delinquency and the most deteriorated housing.

ABCD is not designed as a permanent agency. Its function is to act as catalyst in the development of a broad attack on the social problems of Boston. Congruent with this design, ABCD does not intend to operate programs itself. Its approach is to aid in designing programs, to provide part of the financial support required to get them underway and to conduct research to determine whether or not they achieved their chosen ends.

The problems which ABCD posed for itself can be briefly summarized as follows: Can a set of interventions be designed and implemented in such a way that they appear on theoretical grounds to have some chance of reducing the delinquent behavior of youth in the community and of producing knowledge as to whether or not the interventions did have their desired effects? The reference term which we have used to identify community efforts to achieve those goals is"action-research demonstrations".

Let me try to define briefly some of the distinguishing features of this notion, an "action-research demonstration." If we leave out the adjective action-research, for the moment, and focus on the term demonstration, this term can be used to refer to a broad category of social effort, the principal theme of which is that the programs are designed and conducted on a trial basis and do not purport to represent total solutions to a social problem. Demonstrations are therefore knowledge seeking efforts. However, they vary from those which define success in an administrative sense--was the program workable administratively?--to those which define success in terms of effect--did the program produce the changes it was designed to produce? Demonstrations also vary from those which define "knowing" in terms of the judgments of "experts" to those which define knowing in terms of the outcome of controlled experimentation. And they vary from those where the knowledge seeking effort comes after the action part of the demonstration has terminated to those where it is built into the demonstration from the start. Action-research demonstrations are those which seek knowledge concerning the effects of the program through use of a controlled experimental design which is built-into the total demonstration effort.

Therefore, it is assumed in the following discussion that what we are after are findings from action-research demonstrations. It is also assumed that the minimum criteria for the acceptability of those findings includes:

(a) sufficient information so that the program component is repeatable;

(b) knowledge concerning whether the program produced the effects it was designed to produce; and,

(c) knowledge concerning whether the specific ingredients of the program were in any way necessary to the production of those effects.

All social experimentation is likely to encounter some difficulty with respect to each of the above criterion areas; these difficulties are, I believe, considerably aggravated in the case of community-based, action-research demonstrations, and even more so when they are conducted on a broad scale.

We conceptualized our action-research demonstration project at ABCD in terms of three sets of variables. One is the dependent variable of the project--in ABCD's case, this variable is juvenile delinquency; more specifically defined as law-violating behavior of 12 through 16 year old males residing in specific areas of Boston. The second set of variables are referred to as the intermediate variables. According to the Project's hypothesis, changes in the intermediate variables should produce desired change in the dependent variable. The third set consists of the program variables--the specific interventions by which it is hoped to produce changes in one or more of the intermediate variables.

Therefore, the project has two fundamental, interrelated tasks. One is to find ways to produce the intermediate changes which the hypothesis asserts will be followed by desired changes in its dependent variable. The second is to determine if, when such intermediate changes occur, they are in fact followed by the desired changes in the dependent variable.

These two basic research questions:--(1) did the intervention produce the desired change in the intermediate variable? and (2) were changes in the intermediate variable, if they occurred, related to changes in the dependent variable?--appear to involve fundamentally different methodological difficulties.

The first involves all the difficulties inherent in efforts to implement an experimental design plus the complexities and difficulties imposed by the fact that in ABCD's case an attempt is being made to implement experimental designs in the community. The second question--whether changes in the intermediate variables are related to changes in the dependent variable--involves a number of additional problems, including all the measurement and statistical problems of attempting to relate changes in variables.

As a backdrop against which to explore some of these difficulties, one of the ABCD programs will be described in some detail--the Week-end Ranger Program. A regular summer camp site and its facilities are being used as the setting for a program for already delinquent boys--they must be on probation to be eligible for the program--with the ultimate aim of reducing their subsequent delinquent behavior. In this program, approximately 30 boys leave the community each Friday afternoon and travel about 60 miles on a bus to the camp site. Between then and Sunday evening when they return, they participate in an organized series of activities, including discussion groups, council meetings, work activities and recreational programs.

Briefly, the over-all design is as follows: arrangements were made with the State Probation Commission whereby the local probation offices in parts of the City of Boston provided lists of names of boys who were eligible by reason of age, residence and other criteria for participation in the program. These boys were asked to come to the Probation offices and participate in a study.

Upon their appearance at the office, the boys were pre-tested on several attitude scales--an anomie, an alienation and a values scale. The theoretical tie-in involves the possible relationship between what might be called a disengagement of delinquent boys from the values and institutional system of the dominant society, on the one hand, and their delinquent behavior, on the other. An attempt has been made to build procedures into the program which appear to have some hope of changing the attitudes of these youth and ultimately, according to the model, their on-the-street behavior as well. After pre-testing, the youth were randomly divided into two groups and the members of one group were invited to participate in the week-end program. The members of the other group were designated as ineligible for the program.

Remembering that our aim is not only to know that certain effects were obtained but is also to know with some degree of probability that the effects were substantively related to a particular set of stimuli, one major problem confronting efforts to evaluate programs of this type is the problem of controlling the stimulus. In my opinion, real strides toward the accumulation of definitive knowledge about the effects of programs will not be made until we are able to think through and develop procedures for handling the whole problem of what constitutes the stimulus. The basic question is: What is it that should be repeated if the program appears to work? There are two related but nevertheless operationally separate issues here. One is the design of the stimulus or intervention. The other--and perhaps the more difficult one--is the problem of monitoring of the intervention.

Compared with classical conditioning experiments or even somewhat more intricate experiments such as those involving exposing populations to movies aimed at changing attitudes toward minority groups, the "stimulus" in a program like the Week-end Ranger Program is exceedingly complex. We started out at ABCD with very definite and clear-cut intentions to conduct and evaluate "repeatable" programs. But we admittedly grossly underestimated the difficulties which are involved in both designing and monitoring programs with the goal of repeatability in mind. It is becoming clear to us that this problem cannot be satisfactorily resolved by simply reducing it to the problem of spelling out procedures in great and specific detail as difficult as even that may be. The direction in which we appear to be heading in our efforts to deal with this problem and (needless to say, we are nowhere near to solving it) is toward the development of principles rather than procedures. What this has pushed us toward is the notion of what I call an "impact model,"--a set of theoretical concepts or ideas which trace the dynamics of how it is expected that the program will have the desired effects; a theory which logically interrelates a set of principles and procedures with desired outcomes. If the impact model is sufficiently worked out, a set of working principles becomes available upon which practitioners can draw not only for the design of programs but also to make practical decisions about day-to-day program situations.

For example, in the Week-end Ranger Program there are different tasks to be performed, the boys must be allocated to work groups in some way, and the problem of the non-worker in the work group must be dealt with. Should the boys choose the task they work at? Should they choose with whom they will work? And, how is the non-worker to be handled? The point I am trying to emphasize is that a satisfactorily developed impact model would logically imply that certain decisions rather than others be made with respect to such problems. Greater adherence to the development of such models should (a) enhance the probability that such programs may have an impact; (b) provide a basis for training program personnel that does not require that a program procedure be specified for every conceivable situation; (c) provide a basis for outside monitoring of the program; (d) provide a rational basis for modifying the program design should it appear that it does not have the desired effects; and (e) provide the basis for a repeatable program which goes well beyond the mere rote repetition of isolated procedures.

On the more positive side, the Week-end Ranger Program illustrates that it is possible to conduct a reasonably well controlled experiment in the community which involves the co-operation of a number of individuals and agencies. We were able to institute even a modified version of randomly allocating subjects to a treatment and a non-treatment group. I say modified because not all of those randomly selected for the experimental group agreed to participate in the program, and therefore the exposed and the unexposed populations do not constitute two truly random samples from the same population. In addition, we were able to obtain the necessary co-operation for rather extensive pre-testing of both experimental and control youth. It is likely that some version of a pre-post test design is going to be necessary in such experiments because of this element of voluntary self-selection to participate on the part of the experimental group. Thus we are eventually going to have to (and because of this co-operation we will be able to) rely on covariance adjustments to bring the experimental and control groups back into line.

It is worth noting, however, that a main reason we were able to get support for the randomization procedures was because of the very limited number of openings in the program. But there is still great public resistance to and considerable lack of understanding about randomization. This problem is likely to be even more serious in the case of really massive programs in which there appears to be room for everybody. This is likely to be particularly true where randomization to non-treatment groups is involved.

Furthermore, in addition to the ever present abhorrence of "denial of service" there is a very strong proclivity on the part of practitioners to believe that they know which type of person will benefit most from a particular program. Therefore, co-operating practitioners designate more people for a program than there are openings only with great reluctance. There is also a related tendency for practitioners to want the most deserving youth to receive the opportunity to participate in special programs. Unfortunately, in the Week-end Ranger Program it is presently impossible to determine the extent to which these two tendencies are operating in the selection of candidates for the program.

There are two basic problems here which relate to potential findings. One is that if only the most deserving are selected--even from among probationers--the possibility of program impact may be lessened because both the experimental and the control subjects may fare very well according to the outcome criterion. The other problem is that when the selection is left to the personal preference of the practitioners the representativeness of the demonstration population relative to some larger population will be unknown.

I think there are several possible lessons here of relevance to the evaluation of anti-poverty programs. One is that it is already clear that the overwhelming pressure is going to be on doing rather than evaluating. This one-sided emphasis is unfortunate. Be that as it may, random allocation to treatment and non-treatment groups is not likely to be frequently possible. But, random allocations to alternative treatments may be. This means, however, if such an approach is to be carried out well, the alternative treatments should be thought through very carefully so that at a minimum they are different and not camouflaged versions of the same basic idea. The impact model--the set of theoretical concepts or ideas which trace the dynamics of how it is expected that the program will have its desired effects--again rears its annoying head, and in turn a hard look at what the goals--the outcome variable --of such programs are and how to measure them will be required. Since even the broad scale anti-poverty programs are not likely to be any better off as regards knowledge of the representativeness of the populations they will be dealing with and they are also going to have to face the problem of self-selection for participation, extensive pre-testing with good instruments is going to be a must if anything resembling definitive findings is to emerge. Not only should there be common use of some of the same instruments across similar programs within communities but also across similar programs between communities. For the first time we might have some cross-community comparative material concerning the populations being reached and the changes being observed.

Just as there are conceptual and methodological problems (some of which have been discussed previously) in attempting to evaluate the effects of programs on their direct outcome variables--what we have called intermediate variables--there are other, perhaps even more perplexing, ones involved in attempting to relate changes in these intermediate variables with the dependent variable of the over-all project.

The ABCD Youth Opportunities Project hypothesis states that certain changes will be followed by certain other changes. The programs are designed to expose members of the target population to procedures which will hopefully produce changes in the individual or his environment. Each of these changes is expected by the hypothesis to produce an increment of improved behavior--less law violation--on the part of the individual. It is the Project's hope that for each program significantly more of the experimentals than the controls will experience the desired change and those experiencing such change, whether they are experimentals or controls, will manifest a reduction in law-violating behavior.

It must be re-emphasized that the

hypothesis asserts a relationship between two sets of changes, not between two static conditions. Using the Weekend Ranger Program as the example again, the hypothesis does not assert that low anomie or alienation scores or high value scores will reduce the law-violating behavior among those manifesting score changes in specified directions.

The problem of obtaining reasonably reliable change measures preceeds the problem of relating change measures, since attempting to relate sets of unreliable change scores does not appear to be too promising a game to play. There has been, of course, a long-standing concern for the problem of the reliability of scores. Interest in the reliability of change scores is somewhat more recent and is receiving increasing attention among statisticians and psychometricians. Problems arising out of the mathematically demonstrated greater unreliability of change scores relative to the reliability of the scores from which they were derived and problems arising out of demonstrated regression to the mean tendencies in test-retest situations are likely to remain central as well as difficult issues for those who are brave or foolish enough to pursue this change problem.

The problem of the relationship between sets of change scores has, to my knowledge, received little consideration in the literature and undoubtedly also involves serious statistical and mathematical difficulties. Measurements of each variable at a minimum of three points in time are required to provide some estimate of the shape of the curves involved. Two of the problems involved are (1) the relationship between the shapes of the curves--the change curve for the intermediate variable and the change curve for the dependent variable--and (2) the question of the time lag throughout the series and between the two sets of changes: When are the presumed effects of the program on the intermediate variable expected to take place? While the program is going on? After participation in the program has terminated? And for how long are the effects supposed to last? How long a time is expected to lapse between the changes in the intermediate variable and their presumed effects on the dependent variable? What are their relative rates of change? These and similar questions are directly related to some very practical issues such as the amount of success a project can possibly have during some specified demonstration period. If there is considerable lag or the rate of change in the dependent variable is relatively low, much of the effects of the demonstration may take place after the cut-off point for the evaluation of the Project. Again the need for a theoretically based impact model is, it seems to me, underscored.

Of the many other problems which beset efforts to conduct and evaluate large-scale action programs, there are two that I would like to bring to your attention in the short time remaining. One is the problem of the meaning of change in the dependent variable--in our case, a reduction in law violating behavior--and the other is the problem which arises from the fact that members of the target population may--in fact, undoubtedly will--get involved with more than one of the programs and that this involvement is non-random.

The first decision we made concerning the definition of change in the dependent variable was that we could not use time comparisons of area rates of delinquency as a basis. ABCD's aims were to change behavior, not to move law-violating people out of an area and non-violating people into it. Therefore an area delinquency rate comparison over time was rejected as a basis for measuring change since wide variations in delinquency rates may occur over time simply because of changes in the constituency of the population. It was decided that a reduction in law violating behavior would have to be measured in terms of the behavior of a specified population--that is, a cohort of individuals.

Another major problem in defining how change in the dependent variable was to be measured is the known relationship between age and delinquency. Beginning around 10 or 11, age-specific delinquency rates increase rather sharply up into and through the late teens. Therefore to simply compare a given individual's behavior at age 15 with his behavior at age 14, 13, 12 and so on would lose sight of the fact that the probability of a delinquent act increases as he gets older. If a cohort of 15 year-olds committed the same number of delinquent acts at age 15 as they did at age 13, for example, this might not look like a reduction--and in terms of absolute numbers it is not--but in terms of what might have been expected of them it is. We arrived at what we felt to be an inescapable conclusion--within the framework of our approach--namely, that a reduction of law violating behavior must be defined in terms of a comparison of an observed measure with an expected measure. That is, a prediction instrument is required to provide an estimate of the law violating behavior which would have occurred had there been no intervention.

I suspect that very similar problems will arise if efforts are made to take a hard look at the possible effects of various components of large scale community efforts to deal with poverty. To take one variable, employability, which is central to most of the poverty proposals I have heard discussed, this dimension or characteristic of individuals is also a function of age. For example, it is quite well known that the great bulk of the very difficult to employ 16 to 21 year-olds begin to disappear into the job market and from the unemployment roles as they approach their middle twenties. Therefore, if evaluations of community programs dealing with this particular segment of the population are based upon observations of their employment history subsequent to exposure to one or more anti-poverty programs, the success observed may be much more apparent than real. What is needed is a measure of their employment status and prospects at some point in time as compared with estimates of what would have probably been the case at that same point in time had there been no intervention.

The last issue that I have time to discuss with you is that of multiple-exposure to programs. This has presented the ABCD Youth Opportunities Project with distinct methodological difficulties. It is likely to be an even greater problem for any effort to evaluate the effects of anti-poverty programs. Two tendencies combine here, I believe, to aggravate the problem. One is the inclination on the part of practitioners to want to shower programs on the members of the target population. The other is the sheer amount of money that is involved and the resulting large number of programs that are likely to be conducted. This is an extremely important issue if we are serious in our desire to ultimately acquire knowledge concerning the most efficient allocation of human and financial resources. For if the members of the target population participate in a number of different programs and even if desired change occurs and is measured, there must be a way devised to sort out the relative contributions of the different programs to the outcome. Otherwise, in order to produce the same results again the whole menagerie of programs would have to be repeated even though only a relatively few of the programs may have actually contributed to the desired outcome. Again, a cohort and a prediction instrument appear to be indispensible to the solution of this problem. Individuals must be grouped according to the programs they have participated in--in our approach, according to the intermediate variable changes they have experienced--and then the groups compared on the differences between observed dependent variable and expected dependent variable behavior.

In summary, I have tried to sketch for you some of what I believe to be the major issues facing action research. Roughly, these issues fall into three categories: (1) general action-research administrative problems; (2) general over-all action-research design problems; and (3) basic methodological problems. I would include in the first category the need for a much more vigorous effort to increase general understanding of the need for this kind of effort and of the function of some of its tools, such as randomization. I would also include here the careful development of a kind of contract format whereby the various parties involved would understand and commit themselves to specified goals, principles and procedures. I would include in the second category the need for the development of impact models and the recognition that action-research is going to require a delicate blending of action ideas,theory and research technique. And, in the third category I would include the need for solutions to the problems of the measurement of and relationships between changes and the development and application of prediction instrument technique.