There is general agreement that human services (i.e., services that depend on direct interpersonal contact between a deliverer and a client) are difficult to evaluate. This article points out some of the sources of this difficulty: first of all, the theories lying behind the delivery systems are often deficient. Second, delivery of such services is highly operator-dependent and hence often radically transformed in delivery, in ways that tend to negate the intended treatment effects. A strategy for the evaluation of human services delivery is proposed consisting of several steps, in which the underlying theory is first tested, next the ability of any system to deliver the services in question, and finally whether given delivery systems can deliver a relatively pure version of the service.

ISSUES IN THE EVALUATION OF HUMAN SERVICES DELIVERY

PETER H. ROSSI
University of Massachusetts, Amherst

The main message of this article is that evaluations of human services delivery systems are difficult to accomplish to the satisfaction of either evaluators or the professionals and administrators responsible for the design and operation of the systems, a theme that can hardly be news to the reader. Going beyond merely the reporting of troubles, this article elaborates this issue in two ways: first, we attempt to provide an understanding of why human services are so hard to evaluate satisfactorily, reviewing in the process both the nature of human services delivery systems and characteristic evaluation approaches. Second, we propose an evaluation strategy that is especially appropriate for human services and which attempts to make possible more satisfactory evaluations.

AUTHOR'S NOTE: *The preparation of this article was aided by a grant from the Russell Sage Foundation, "Measuring the Delivery of Public Services," whose assistance is gratefully acknowledged. The comments of several colleagues, Wayne Alves, Richard Berk, and Huey Chen, on an earlier draft were most helpful in sharpening the article at several points. An earlier version of this article was delivered at a conference on Issues in Service Delivery in Human Service Organizations, held at Wingspread under a grant from the Wood Johnson Foundation and the Silberman Foundation, June 1977.*

EVALUATION QUARTERLY, Vol. 2 No. 4, November 1978

There are many activities that go under the name of evaluation, ranging from offhand opinions, through news reporters' haphazard investigations, to social science research efforts as rigorous as the current state of the art in basic social science permits. For the purposes of this article, I want to restrict the term evaluation to the application of current, state-of-the-art social science research methods to the assessment of whether given social policies can achieve or are achieving their intended aims. This restriction excludes those evaluating activities that do not pretend to be social science and those that pretend but do not succeed. Admittedly, such characterizations are judgment calls, but ones upon which I am sure we would mainly all agree. Also excluded are researches that do not attempt to assess whether programs are fulfilling their goals, and hence are purely descriptive accounts. Policy analyses are also ignored on the grounds that they do not involve primary research activities.[1]

As we shall see in a later section of this article, this definition of evaluation research does cover a very wide range of activities, including research that attempts to discern what are the goals of a program, monitoring activities that seek to ascertain how a program is operating, process or formative evaluative activities, as well as impact assessments and field experiments (see Rossi and Wright, 1977).

If there is any empirical law that is emerging from the past decade of widespread evaluation research activities, it is that the expected value for any measured effect of a social program is zero. In short, most programs, when properly evaluated, turn out to be ineffective or at best marginally accomplishing their set aims. There are enough exceptions to prevent this empirical generalization from being phrased as the "Iron Law of Social Program Evaluation," but the tendency is strong enough to warrant placing bets on negative evaluation outcomes in the expectation of making a steady but modest side income.

The disappointments that have arisen from the results of program evaluations have led, on the one hand, to a reconsideration of programs and, on the other hand, to a reconsideration of evaluation research as an activity. Neither reassessments has led to much progress up to this point. It is apparently the case that evaluation research does not lead immediately to radical improvements in social programs. Nor does the failure of evaluation research to find positive effects of programs lead to its being discarded as an approach. Indeed, one can make the case that nothing succeeds like failure, a paradox whose resolution rests on the understanding that it is composed of half truths. Thus, some believe

that evaluation results reflect reality while there is nothing wrong with evaluation research methods. Others believe the exact opposite. Each camp partially neutralizes the other with the result that there is widespread skepticism both about social programs and about evaluation research. Yet so far we cannot do without either.

ON THE NATURE OF HUMAN SERVICES DELIVERY SYSTEMS

The tertiary sector of our economy has been growing at a faster pace than any of the other sectors. Our affluent society apparently has mastered the problems involved in primary extraction and in the manufacture of finished products and has turned of late to putting the finishing touches on the "quality of life." This fine tuning of our social machinery involves the extension of existing human services and the invention of new forms of such services. Through government agencies and private entrepreneurs we now channel a considerable portion of our GNP into providing services that depend essentially on the delivery outside market mechanisms of some sort of "product" to clients through the use of more or less trained intermediaries. The essential aspect of human services is that the mode and content of the delivery itself is the product that is being delivered. Thus education is the interaction between students and the schools, the primary aspect of which is the activity of the classroom. Similarly, job counseling is the contact between a counselor and a client and the product is the content of these encounters. Of course, there is more to human services than human interaction in face-to-face encounters: schools consist of physical structures in which the schooling takes place, textbooks, writing implements, audio-visual aids, and the like. Similarly, job counseling may involve the use of aptitude tests, pamphlets, audio-visual displays, and so on.

The essential premise of human services systems is that there are pockets of deficiencies in our social structure that can be corrected through such encounters, or that naturally occurring processes accompanying human development can be speeded up or made more efficient with the use of human services delivery. We know that without formal schooling children will grow into adults and acquire some degree of verbal language skills. We also know that some families alone could rear children who are literate and have basic mathematical and other

skills. But, the family as an institution is not very good at imparting such skills: without the public educational system, the disparities among families in these respects would tend to exacerbate inequalities in skills and knowledge while at the same time lowering the average levels of skills and knowledge below those seemingly required by our need for a relatively literate and knowledgeable labor force. Public education is simply more efficient than the family in aiding young children to develop the necessary skills and knowledge.

What is apparently very clear with respect to education, is not as clear in the cases of other types of human services. Thus, we do not have any institutions that are universally found in all societies that are concerned with the detection of crime or with the rehabilitation of prisoners that are released to freedom. Nor do we expect that unemployed persons will have on hand all the skills that would make it easy for them to get other jobs or that their families and friends will have all the information on hand that will make the transition back into the labor force as easy as possible.

Furthermore we recognize that there are gross inequalities among individuals and households in their abilities to deal successfully with the world about them. Although money cannot buy everything, the rich and the powerfully connected can buy many things on the open market that make life easier. The poor, those disadvantaged in some respect, or those whose skills and aptitudes are deficiently below normal all apparently need help or they will sink further down to new lows in depravity. Obviously, here is where human services come to the rescue.

It should be noted that there are two aspects to these problems of deficiencies in individuals and households. On the one hand, a deficient human being or household lacks or is deprived of experiences and/or resources that are essential to adequate functioning. Thus, according to this view everyone should have a chance at a reasonable job, or a reasonable chance to recover from an illness or injury. In short, we hold to concepts of *social minima* for units of our society. On the other hand, a society with deficient individuals and households suffers some disabilities because of those pockets of deficiency. An unrehabilitated released felon will commit additional crimes. An unemployed man is not contributing to the GNP. There is an underlying concept of a *societal minimum*, a minimum level of functioning for the society.

Corresponding to these two aspects of the problem of deficiencies in individuals and households, there is a duality that, on the one hand, expresses concern for individual suffering under deprivations, and, on

the other hand, concern with social control. Thus an unrehabilitated prisoner is deemed to be a potentially unhappy person and also someone who is a menace to social order. In this conception of deficiencies, both society and the deficient individual or household have at least parallel if not identical interests: an unemployed person wants a job and the society wants to have a low level of unemployment or underemployment. Fix one and you fix the other. It should be noted that there are some areas where the parallelism is not so obvious. Certainly, society may want a low crime rate but some professional criminals might have little interest in being rehabilitated to follow occupations that are less interesting and remunerative. This duality becomes expressed with particular force in some human services delivery systems where the interests of clients and the interests of society diverge sharply.[2] We will return to this theme again in this article.

The establishment of a human services delivery system rests upon a number of critical assumptions, as follows.

(1) There are deficient individuals, households, or institutional arrangements. These deficiencies prevent optimal functioning of some individuals and households. Furthermore, the presence of these individual and/or institutional deficiencies in the society presents problems to the society.
(2) If the deficiencies can be corrected or compensated for, functioning can be changed so that individuals and households can function "normally" through the use of some sort of human service "treatment."
(3) The human services "treatment" can be delivered uniformly and widely through the training of delivery personnel and through the placement of them in an organization.
(4) There are no serious conflicts of interests between the social control goals of human services and the goals of clients.

The evaluation of human services delivery systems ordinarily takes place around points two and three. There is little doubt in conventional establishmentarian views that there are deficient institutions, individuals, and households in our society.

There is also agreement that it is possible to compensate for or correct these deficiencies by the provision of some sort of services. Thus schools can be viewed as supplementing family units in the provision of socialization, police as providing social control services that neighborhoods cannot otherwise provide. Transfer payments make up for deficiencies in household income, training programs for deficiencies in skills. Rehabilitation programs attempt to remedy the defects of previous socialization, and so on.

The conventional establishment viewpoint does not go unchallenged, however, by alternative models. Conservatives may recognize the deficiencies but deny that human services can provide effective remedies, or assert that the remedies create more problems than they cure. From the left come diagnoses that assert that deficiencies are inherent in the main structures of our society and can only be remedied by radical restructuring of the society itself. In between are viewpoints that stress institutional deficiencies and those that stress individual or household deficiencies. The extreme conservative and the extreme radical viewpoints, however, do not find their way into evaluative activities since they are based on models that contradict the policies that lead to the provision of services.

Little or no attention has been paid to designing or carrying through evaluations that question the fourth assumption on the above list. The utilitarian heritage of our liberal social philosophy equates—at least in the long run—the utility for an individual with social utility. However, in the case of some social programs this identity of interests is not clear. For some persons of low levels of educational attainment and correspondingly low repertories of job skills, getting a job may not be better than remaining on welfare, as many of the welfare mothers enrolled in the WIN program have learned. The WIN program that attempted to move welfare mothers into the labor force assumed that a low-paying job would appear more attractive to mothers of young children than remaining on the welfare rolls.

The main thrusts of evaluation center around assumptions two and three. It is assumed that there is some way in which these deficiencies in institutions, individuals, and households can be corrected (or compensated for) through treatments and that such treatments can be delivered effectively to clients with reasonable cost-to-benefit ratios. While few evaluations center around the fourth assumption, it can be shown that this assumption when violated in fact plays an important role in the failure of human services delivery systems.

THE FAILURE OF TREATMENT

Assuming that pockets of deficiency exist, then whether a treatment can be devised that will reduce the size of such pockets depends clearly on whether or not the conditions generating the deficiencies are properly

understood. In short, treatment depends on the existence of a valid model of how the deficiency is produced and/or maintained. Thus, if a model of black unemployment differentials sees the high unemployment rate among blacks as due to a lack of skills that match current labor market demand, then a reasonable treatment to apply is vocational training. Obviously, if the model is not valid, then vocational training will fail as a treatment. Or, if the model underlying the design of a prison is that criminals are members of a deviant subculture, then the design might minimize contact among prisoners and emphasize lofty sermons on the straight life. Clearly such a treatment is likely to fail.

An infinity of models may be devised to explain such deficiencies as poverty, crime, unemployment, illiteracy, and the corresponding treatments may also be infinite in number. Indeed, the same model may lead to quite different treatments depending on whether one emphasizes one or another part of the model. Thus a production-function model for public education might lead one to emphasize school inputs to learning or emphasize family inputs, the first treatment leading to heavier investment in teachers, schools, or teaching methods, while the second leading possibly to childrearing instruction for parents.

Furthermore, every treatment can be shown to "work" in the sense that, after experiencing the treatment, some of the deficient persons or households will improve. Our society is highly stochastic, with individuals and households moving from state to state with probabilities that are significantly large. Thus, poor individuals and households are quite likely to cross the poverty line if left untreated and some portion of those who are treated will also cross the line and appear to have been affected by the treatment. While the old are not going to get young over time, the opposite process appears to make young adult criminals into more law-abiding older adults. Men and women who never finished high school on time often manage to get their diplomas later in life. Some persons who have received vocational training will have higher wage rates after training, but so will their counterparts who have not taken training but who are just a few months older. In short, "spontaneous remission" is characteristic of these deficiencies.

Finally, it is difficult to separate the treatment from the manner of delivery. Human services treatments that are given by exceptionally devoted persons are as likely to be efficacious because of the devotion expressed in the delivery as because of the treatment. Individual tutoring of children seems always to work for the rich who have been able to hire skilled and devoted teachers. The results of any brand of therapy

wielded by a devoted psychotherapist is probably as good as any other. However, the same treatment administered by the indifferent and unskilled may fail to have any impact. The *delivery* of human services treatments is critically important to their effectiveness, a theme that is treated at length in the next section of this paper. Before proceeding, however, it is important to point out that treatments can fail in the first place because they are inappropriate to the problem, because they have been generated by an invalid model of the phenomenon in question. They can also fail because the treatment itself has been poorly specified and is in fact indistinguishable from its mode of delivery.

THE FAILURE OF DELIVERY SYSTEMS

Assuming for the moment that an efficacious and theoretically valid treatment has been devised, then the next question is whether it can be delivered on a large scale. A related problem is whether the delivery system devised for the treatment either negates the treatment or transform it into something else. The delivery system is of special importance in human services because the ultimate delivery point is a human being whose needs may be such that they work at cross purposes to appropriate delivery.

A few examples of how delivery systems can fail may be appropriate at this point.

(1) The problem of the nonprogram. This is the case where a delivery system has been set up or an existing system has been designated as a deliverer, but no treatments are delivered. Lest the reader believe that these are rare instances, it turns out that there are many examples, as follows.

A network of advisors was set up to provide advice to new M.D.'s sent out to rural areas to help the doctors become accustomed to their new (and presumably strange) environments. Evaluators sent out after a year or so of the program discovered that the advisors rarely contacted the doctors after the first visit. Apparently, advisors (who kept on receiving their stipends) discovered, as did the doctors, that no advice was needed or wanted or appropriate.[3]

The Office of Education's attempt to find out exactly what new educational services were delivered to children in schools in poor

neighborhoods under Title I of ESEA have been repeatedly frustrated by the inability of local educational authorities to describe their Title I activities in any detail (McLaughlin, 1975).

Attempts to decriminalize alcoholism and to use the police to bring alcoholics into treatment centers in Washington, D.C. and Minneapolis have found that when the police stopped arresting people for public drunkenness they did not necessarily scoop them up and bring them to treatment centers. Police get credits for arrests but not for ambulance service.

(2) The problem of creaming. Although the world is stochastic, it is also lawful. Hence a delivery system can simulate success by delivering treatments to individuals who are most likely to recover or to households that are most likely to rise spontaneously out of their deficient state. Some examples follow.

In the first years of the Job Corps, the screening methods used to test for "poverty" eliminated those whose families were too affluent, and for "potential," eliminated those who seemed "unlikely to benefit from the treatment."

FHA guarantees for mortgages, originally proposed as a means of helping the poor to purchase homes, became a subsidy for the middle class as FHA administrators took care to guarantee loans only for those who had good credit ratings.

Fellowships for graduate study, proposed as a move to bring more talent into particular fields of national importance, are given out competitively, assuring that those are subsidized who would go into such fields anyhow because of interest and aptitude (Davis, 1962).

A Planned Parenthood Clinic set up in the early 1950s on the South Side of Chicago ostensibly to provide contraceptive services to black ghetto residents found itself so swamped with student clients from the University of Chicago that it made few efforts to reach the blacks of the South Side.

(3) The problem of delivery negating treatment. The modes of delivery may operate in ways that negate the treatments. Some examples follow.

Case workers in a state public welfare system were found to have developed a classification of "good" clients and "bad" clients, the former to whom they offered all the options for payments allowable

under regulations and the latter to whom they gave payment options only when asked for specifically. "Good" clients were those who presented themselves as suppliants and expressed gratitude easily for help proferred while "bad" clients were those who demanded payments as a matter of right.[4]

A negative income tax experiment, sparked in part by a desire to test a system that did not have a demeaning means test, developed a system of monthly family income reports that kept closer track of participating family earnings than could any public welfare system (Rossi and Lyall, 1976).

An experiment that was to test the effectiveness of group counseling in prison used prison guards as group leaders (Kassebaum et al., 1971).

There is some evidence that the contract learning experiments were sabotaged by the school systems into which they were conducted, with the consequence that the treatment was delivered in only a subset of the thirteen schools originally contracted for (Gramlich and Koshel, 1975).

(4) The problem of uncontrolled treatment variation. Discretion on program implementation left to the front line delivery system may be so great that treatments vary in significant ways from site to site. Such intersite variation may exist especially when treatments are forms of delivery with the content of services to be delivered left to local delivery systems to determine. Some of the best examples can be found in the early programs of the Office of Economic Opportunity, as follows.

The Community Action Program left considerable discretion to local communities to engage in a variety of actions, constrained only by the requirement that there be "maximum feasible participation" on the part of poor citizens. As a consequence, it is almost impossible to document what the CAP programs in fact did.

A similar lack of content definition also characterized the Model Cities Program.

Project Head Start gave money to local communities to set up preschool teaching projects for underprivileged children. Programs started up had a variety of sponsoring agencies, differing coverages, varying content, and the like. To evaluate Head Start is to evaluate a program that is so heterogeneous in essential respects that it cannot be called *a* program at all.

(5) The problem of ritual compliance. Lack of commitment to a program on the part of a front line delivery system can have the result that minimal delivery of the program occurs. The treatment in not negated, it simply is watered down almost to the point of nonexistence.

In an effort to assure more contact between professors and their students, a state legislature mandated semester reports from each professor with detailed counts of "contact hours" to be entered. A considerable professorial effort went into stretching every potential contact opportunity into a contact hour. Thus, there were more advisees reported than there were students to be advised.

To comply with affirmative action directives, university departments often place ads in national professional publications for positions already informally filled, announcing the selection officially only when replies to ads have been received and the resulting applications rejected.

(6) The problem of overly sophisticated treatments. Some treatments that might work well in the hands of highly trained and highly motivated deliverers may be given to a mass delivery system whose levels of training and motivation are considerably less and hence the treatment fails. In short, there is a considerable difference between pilot and production runs of sophisticated treatments.

Thus, although many educators have come forth with teaching methods that have worked well within their experimental classrooms and schools, the adoption of such teaching methods in ordinary school systems have not proved very successful. (In part this is a problem of the delivery being part of the treatment.) Computer-assisted learning, individualized instruction, and so on, are examples of techniques that seem to do less well when applied outside of the centers where they were developed.

(7) The problem of client heterogeneity. A treatment that works well with one type of client may not work well with another. This problem is especially acute if the pilot tests of a treatment are done with a special population and then applied in a production run with a quite different client population.

In the New Jersey-Pennsylvania Income Maintenance Experiment, ethnicity turned out to be one of the characteristics distinguishing subgroups with different work effort responses: black households increased their work effort under a guaranteed income plan; whites

decreased; and Puerto Ricans showed no significant work effort effect (all compared with controls who were not on payment plans)[Rossi and Lyall, 1976].

The early somewhat spectacular finding that preschool children could be taught very effectively using electric typewriters, did not hold up in repeated studies. Apparently middle-class preschool children were aided by the method, especially in the skilled hands of its originator, while lower-class children were unable to benefit similarly from the treatment.

(8) The problem of client rejection of treatment. This problem stems from rejection of the treatment by potential clients with the result that the treatment cannot be delivered to the extent desired.

The Housing Allowance Experiments currently underway have experienced participation rates considerably below (30%-40%) full coverage of eligible population groups (Carlson and Heinberg, 1977) despite apparently obvious advantages to potential participants.

Community Mental Health Centers designs to provide outpatient treatment to clients in need find that it is difficult to get potential clients to come to the centers. Even patients conditionally discharged from state mental hospitals who have been assigned to centers as a condition of their discharge often do not appear at centers for their treatments.

WHY TREATMENTS FAIL

The litany of delivery problems outlined above has its roots in the fact that human services delivery cannot be made operator-free. Rules for deliverers can be developed that seemingly take discretion out of the hands of operators, but the proliferation of rules itself can be seen as one of the sources of operator discretion. Thus the manuals governing (in principle) the activities of a caseworker in the Massachusetts Public Welfare Department are nearly a foot thick, more than anyone can be expected to know intimately. Hence, there is considerable variation from caseworker to caseworker, from local office to local office, and from season to season as to which rules are enforced and which provisions of the manual are in fact used.

Another way of putting this problem is that many human services treatments are insufficiently robust and are unable to survive mis-

handling on the part of delivery systems. Even seemingly robust treatments in the form of transfer payments can become transformed in the course of being administered by a delivery system.

Often insufficient attention is paid to the problem of motivating human services operators to deliver treatments as specified. Thus, in the case of the Washington, D.C. and Minneapolis decriminalization of public drunkenness, no thought was given to motivating the police on the beat to escort drunks to the alcoholism treatment centers. Similarly, the contract learning experiments made not attempts to reassure the regular teachers who feared that their jobs were threatened by the private contractors who competed with their regular classes.

Another source of difficulty may lie in the professionalization of some of the deliverers of human services. It is of the essence of a professional occupation that incumbents function with minimal supervision, the assumption being that professionals need little supervision because their training fits them to make appropriate discretionary decisions about the content, pacing, and outcome of their work. When to professionalization is added immunity from market and price effects, than a delivery system may be particularly difficult to affect by administrative directives mandating changes in delivery practices. Thus, my colleagues and I found in a comparative study of fifteen major metropolitan areas that police practices and public welfare agency practices were more subject to variation from place to place than were the practices of educators (Rossi et al., 1974). Indeed, it was possible to predict more closely how police behaved toward black residents (as reported by blacks themselves) on the basis of policies professed by police chiefs and mayors than it was to predict how teachers behaved toward their pupils on the basis of pronouncements of mayors and school superintendents. The behavior of the caseworkers in public welfare agencies fell in between but resembled more the case of the police than the case of public education.

The current issues surrounding the cost of medical care and its quality also illustrates how difficult it is to establish some modicum of control over a highly professionalized delivery system. Despite the proliferation of hospital planning councils, hospitals still tend to build more beds than they need and to install expensive equipment the use of which frequently fails to justify the investment. Serious abuses exist in the overuse of surgery and in the wholesale prescribing of tranquilizing drugs and antibiotics in cases where such treatments are clearly not

indicated. A peer review system functions only when the abuses are flagrantly obvious.

These observations suggest that we need an engineering counterpart to the "pure" social sciences. An academic mechanical physicist can design a bridge according to a new concept, but it takes an engineer to select the materials, prepare sites, and work out details of how the new design should be implemented. Insufficient attention has been paid to the development side of "research and development" activities in the social sciences.[5] We need to devise ways in which we can test out various production forms of a treatment in which the characteristics of delivery system are taken into account and to develop treatments robust enough to survive considerable mishandling.

A STRATEGY FOR EVALUATING HUMAN SERVICES DELIVERY SYSTEMS

In this section we propose a strategy for evaluating human services delivery systems that takes into account the characteristics of such systems as described in previous sections. Before doing so, however, it is important to point out that only those human services delivery systems (or any social program, for that matter) can be evaluated whose intended aims are delimitable, measurable, and not inherently contradictory. For example, a program that is designed to increase the quality of life in America cannot be evaluated until specific content is given to the phrase "quality of life," a difficult, if not impossible task. Incompatible goals are also a contradiction of evaluability: thus a preschool educational program that is designed to serve all segments of the class structure and at the same time decrease the gap in learning between classes is most likely contradictory in aims and hence cannot be evaluated.[6]

Assuming human services programs that have definite, noncontradictory, and measurable goals, then there are three points at which the treatment involved can be and should be evaluated.

(1) The question may be raised whether the treatment *is effective* in achieving its goals, given the most favorable delivery method.
(2) There is the question whether the treatment *can be delivered* by a delivery system that can reach the appropriate target population at reasonable cost levels while maintaining the integrity of the treatment.

(3) There is the issue of whether a given delivery system that, in principle, can deliver a treatment *in fact will do so* at a level of quantity and quality necessary to assure a reasonable level of effectiveness and will be accepted by the target population.

Note that as far as evaluation outcomes are concerned, these three questions are interlocking ones. A treatment that has been found to be ineffective on the first level cannot be evaluated as successful on the second and third levels. Similarly, a treatment that is judged effective on some pilot basis may fail to be effective in practice because it cannot be delivered either by the best of all possible delivery systems or by the usual mass delivery system that an enacted program would use.

This nested quality of the three evaluation questions strongly suggests that an effective evaluation strategy ought to be based on a progression of evaluative activities proceeding from an attempt to answer positively the first question, and so on through all three.

(1) Is a treatment effective? A treatment in principle is effective to the extent that the treatment flows from a model of the phenomenon in question that is a valid reflection of the process involved. Thus a treatment for juvenile delinquency that is based on a model of delinquency in which brain injury is the causative agent is likely to fail because the underlying model is faulty.[7] An effective treatment should be effective, at minimum, under delivery circumstances that are most favorable to effectiveness and, at maximum, be effective no matter what the form of delivery.

These considerations lead to a first step in an effective strategy of evaluation, that is, one concerned with testing the effectiveness of a treatment under maximum favorable conditions or under a varying set of conditions that is sufficiently diverse to make possible the separation of effectiveness of a treatment from its mode of delivery. Indeed, the evaluative activity that is variously called "process evaluation" or "formative evaluation" is often most appropriate to this task.

It also makes sense that treatment evaluations at this stage should be small scale, "pilot" studies that maximize internal validity—in this case, the ability to make strong statements about the effectiveness of the treatment. Carefully designed randomized experiments would be particularly appropriate, assuming that the treatments lend themselves to laboratory or field experimentation, an issue to which we will turn later. If the treatment is one that can vary in amount of intensity and can be delivered by a number of techniques that appear a priori to be

roughly equally appropriate, then a pilot experiment could be designed that would vary level of treatment and delivery technique simultaneously. The end result of such an elaborate pilot phase would be more useful knowledge about the more appropriate levels of treatment and the most effective modality of delivery.

A good example of the type of design suggested above can be found in the Kassebaum et al. (1971) study of group counseling with prison inmates in which several levels of treatment and technique were integrated into a randomized design. One may reasonably raise the question concerning this study of whether the treatments varied enough in levels from one experimental condition to another and whether there was sufficient variation in the techniques of delivery,[8] but the general outline of the design remains a good one and particularly appropriate to the issue addressed in this section.

It should be noted that randomized controlled experiments in human services delivery are relatively rare. The major field experiments in social programs of the past decade typically center around the delivery of transfer payments as treatments rather than human services as typically defined. The five negative income tax experiments involve payments to poor and near-poor families that are conditioned upon their earnings. The current experiments on housing allowances for poor and near-poor families also use transfer payments as treatments, in some experimental groups being conditional upon improvements in their housing. The health insurance experiment currently under way also involves federal subsidies on a sliding scale for full coverage medical and hospital insurance. Finally the Department of Labor financed experiment providing unemployment compensation payments to felons released from state prisons is another example of the use of money as a treatment.

Of course, it is an oversimplification to regard the payments in such experiments as fully comparing the treatments: In fact, the treatments consist not only of the payments but of all the contacts between the paying organizations and families in the experimental groups. In the negative income tax experiments monthly reports of earnings were required as a condition of eligibility and some discretionary powers were given to persons who computed payments.

Nevertheless, it is fair to say that the reason transfer payments are at the core of the treatments studied in the major field experiments is because payments appear to be robust treatments that can be standardized and delivered in relatively fixed ways compared to such treatments

as parole supervision, job counseling, and the like. Such experiments are easier to interpret since one can be more certain that the treatments were delivered—checks can be traced, amounts can be ascertained. In contrast, whether parole supervision of any sort actually took place is problematic and parole supervision can range in intensity from brief superficial contacts between a parole officer and a parolee to more intensive encounters. It is instructive to note that the one negative income tax experiment in Gary, Indiana, that tried to introduce social services as an additional treatment in one of the experimental conditions failed to implement that treatment. It was simply too difficult to standardize sufficiently the social services rendered, to deliver the services in a systematic way, and to get client acceptance of such services.

This suggests that designing human services delivery experiments according to classical randomized design will be very difficult, a task that is sure to tax the ingenuity of experimental design experts and social service professionals. For, unless the treatment can be made more or less standard and delivered in a standard way, then the interpretation of experimental results will be difficult, if not impossible.

(2) Can the treatment be delivered? A treatment that survives the tests suggested above using a randomized design next has to be considered from the point of view of an appropriate delivery system. If the delivery techniques have been varied in the pilot experimental phase, some knowledge about effective delivery has also resulted from this phase. Such results, however, are not to be trusted entirely. A treatment that works well within an experimental context with personnel specially trained by an advocate of the treatment may fail in the field when an attempt is made to use personnel and organizations that cannot match the dedication and skill of the group that has run the randomized experiment. In short, the next task is to test the external validity of a treatment, its ability to be transferred into the "real" world of existing organizations.

There are a few precedents for such delivery system testing that can be cited. The Follow Through Planned Variation evaluations of the Office of Education (Cline, 1975) were intended to perform this function for a variety of compensatory learning techniques, but the execution was flawed. The idea behind the evaluation was to get several school systems each to choose one of several teaching methods, to implement those innovations within the school systems, making provision for reasonable controls within schools. Since there were to be several school

systems testing each method, it was hoped that information would be generated on the relative effectiveness of the different teaching methods and on the relative effectiveness of variations in the delivery systems. The failure of the evaluation occurred because treatments were varied in unsystematic ways when implemented.

A second example is the so-called "Administrative Experiment" designed to test the ability of local authorities to administer a housing allowance program (Carlson and Heinberg, 1977). Local communities were asked to bid for designation as a demonstration for a housing allowance program in their cities. Eight successful bidders were chosen from among competing cities. Agencies within the cities varied from place to place—in some cities the program was administered by separately established agencies, in other cities by an already established housing authority or planning department. Cities were chosen to represent a spread in size and region, although none of the very large metropoli were among the group. Unfortunately, the administrative demonstrations were not monitored carefully enough and with sufficient attention to problems of valid inferences about relative effectiveness so that at the present time it is not possible to make statements about how effective the delivery systems were.[9]

A third example is the Transitional Aid Research Project of the Department of Labor. In a pilot randomized experiment in Baltimore, the U.S. Department of Labor (1977) found that in providing payments resembling unemployment compensation payments to prisoners released from the Maryland state prisons, those who received payments were less often arrested for property crimes in the year following their release. The pilot experiment was run by a dedicated researcher, Dr. Kenneth Lenihan, who recruited a staff of counselors, payment clerks, and so on. Currently the Department of Labor is funding two additional large scale randomized experiments in the states of Georgia and Texas, in which state agencies administer the payment plans and collect data on released felons. Up to the date of writing, the experimental design appears to have been implemented correctly and payment systems are operating well. The purpose of the larger scale experiment was both to replicate the Baltimore experiences of Lenihan and also to test whether or not existing state agencies can administer such payments in a way that would retain their effectiveness. The administration of the plan and the collection of research data on the released felons are being monitored carefully by the Department of Labor and a set of subcontractors.

This Project provides an excellent example of research designed to test whether existing delivery systems can function effectively in delivering a treatment that is known to be effective. One might have wanted a few more replications, possibly ones administered by different state agencies and covering some of the largest metropolitan areas, but the appetites of researchers are well known to be insatiable.

(3) Is a treatment being delivered? Assuming that a treatment passes with flying colors the tests described in 1 and 2 above, there is still the question whether, when implemented as a statutory program with widespread coverage, treatments are in fact being delivered in appropriate ways. To answer this question requires the setting up of monitoring systems that measure and assess treatment delivery.

To the extent that the human services treatment delivered is some interpersonal transaction between a deliverer and a client, the measurement of delivery is rendered extremely difficult and expensive. To the extent that there is some observable, relatively objective outcome of the delivery, then the task of monitoring becomes that much easier and less expensive. For example, it is possible to obtain fairly accurate counts of how many clients have been served by a family planning agency and how many intrauterine devices or other types of contraceptive methods have been prescribed. Similarly, AFDC client loads can be counted, authorized payments summed and averaged, and other quantitative indices defined and computed. What is difficult to measure is the style and content of contraceptive advice given in client visits or whatever counseling takes place in the caseworkers' contacts with AFDC applicants. Were clients treated with due regard for their human dignity? Was the advice appropriate to the client? Did the counseling given resemble closely enough what the program designers had intended to be given? These essentially qualitative aspects of client-deliverer contacts are difficult to measure at an acceptable level of cost.

The remarks made above are not to be taken as meaning that quantitative measures of client-deliverer contacts are not important. Indeed, one can learn a great deal about how human services are being delivered by considering such relatively simple indices as client loads, socioeconomic composition of client populations, counts of specific service delivered, and so on. For example, in the routine monitoring of hospitalizations in a New England state, it was discovered that in one of the hospitals an extraordinary number of appendectomies were being

conducted. Further investigation brought to light the fact that one surgeon was contributing almost all of the surplus appendectomies. The inquiry led to the setting up of a hospital peer review committee that subsequently disciplined the offending doctor. Or, a comparison across states in the per capita state prison populations brings to light some startling interstate differences in criminal justice systems, some states moving prisoners quickly through their systems and others retaining prisoners for longer terms. Such "epidemiological" studies of the functioning of delivery systems can be very valuable for understanding the gross features of the delivery system, for pinpointing problems in functioning, and, in some cases, for laying the basis for an evaluation of effectiveness.[10]

An accounting system that is run by the delivery system itself is clearly the least expensive way of monitoring, although subject to the possibility of generating self-servicing statistics. A reporting system that is useful both to the delivery system and to outside monitors is obviously desirable since such a system tends to motivate the deliverers to maintain high quality. Thus, in the juvenile court system of Connecticut a reporting system was devised that served both as the data base for monitoring operation and as the case file on each juvenile brought through the system. The forms used were developed through extensive consultation between a central research and evaluation staff, caseworkers, and the juvenile courts. As a consequence, the quality of the resulting data appears to be high.

Monitoring the more qualitative aspects of human services delivery is a more difficult and expensive task. Yet there are some good examples: Reiss (1971) placed observers in police patrol cars who filled out systematic reports of each encounter between the police and citizens in a sample of duty tours. In a now classic study of a state employment service, Blau (1955) sat in as an observer as clients were interviewed as they registered in the agency.

"Windshield" surveys have been devised to measure the cleanliness of streets in various neighborhoods as a measure of the effectiveness of street cleaning and garbage removal crews. In some of the studies, streets were compared against "standardized" photographs indicating extremely, moderately, and poorly cleaned streets and the streets rated according to their resemblances to the standard photographs.

Considerable effort has gone into the measurement of human services delivery systems through interviews with clients (or potential clients). Thus, my colleagues and I analyzed interviews with samples

of black residents in fifteen major metropolitan areas that asked about instances of police brutality either experienced directly or known about, as well as satisfaction with the services of neighborhood stores, schools, and public welfare offices (Rossi et al., 1974). In the New Jersey-Pennsylvania Income Maintenance Experiment, participants were asked about their knowledge of the payment plans they had experienced in an effort to discern whether correct knowledge about the plans experienced affected their labor force responses (Rossi and Lyall, 1976).

Direct observation of deliverer-client contacts are obviously expensive and in addition unwieldly as a research operation. It is also not clear the extent to which the presence of observers affects the ways in which human services are delivered. Reiss (1971) informs the readers of his monograph that the police in the patrol cars soon became accustomed to having observers around, but this observation can only be an impression.

Client interviews are cheaper and are potentially quite useful. It is important that such interviews not rely simply on global assessments of delivery system behavior (e.g., how satisfied are you with your caseworker?), but also provide quite specific information on the content and utility of contacts. For example, it probably is more useful to know whether the deliverers address clients by their first or last name than it is to know the clients' assessment of how politely they have been handled. Similarly, it is more important to know whether a policeman stopped and frisked an arrestee than it is to know whether he thinks the police "are doing a good job."

This article has devoted so much space to the topic of monitoring ongoing programs out of the conviction that such activities are extremely important in the assessment of the effectiveness of human services delivery. A treatment that is not being delivered or is being delivered in a defective way obviously cannot be effective, although correct delivery is not any guarantee of effectiveness. The same ingenuity that has brought social science research to its present state of competence in other areas, if focused on the problem of program monitoring, should result in effective and informative monitoring operations. A monitoring system is useful not only for evaluation but also for correcting administrative faults. A human services systems administrator who does not know whether his program is operating as designed is obviously an inefficient administrator who has to operate largely in the dark.

(4) Is the production run of a program effective? The final question is whether a treatment that has been proven effective in a tightly designed pilot experiment, and has been shown to be delivered correctly and efficiently by a delivery system, is in fact having its intended effects when implemented as a matter of social policy. Presumably, a treatment that has survived the previous hurdles should be effective, but not necessarily so. There are many intervening events that can lead to ineffectiveness as an enacted social policy. Specification error (or erroneous models) in the original experiment may have misled the investigator into mistaking a correlated effect for a real one. Historical shifts may have made an appropriate model into an inappropriate one; for example, the unemployed in times of high unemployment may contain a different mix of population types than does the unemployed in times of low unemployment. Or women seeking birth control information in a period of high fertility may be quite different with different needs for treatment than women who come to birth control clinics in a period of low fertility. It is also possible that the pilot experiment inadvertently creamed the target population of clients.

An ongoing social program that is already in place and functioning at its intended coverage and funding cannot be evaluated through the use of the more powerful research designs. In particular, randomized experiments ordinarily cannot be used since the construction of a control group through randomization will mean depriving some individuals or households of treatments to which they would otherwise be entitled by law or ethics. Hence, such programs usually can only be evaluated by quasi-experimental methods. It should be pointed out that the success of quasi-experimental methods depends very heavily for their utility on a valid understanding of the causal processes underlying the phenomenon in question. Thus, if we want to evaluate the effectiveness of family planning programs in reducing fertility rates, we clearly have to know something about what effects fertility in order to hold constant in our statistical models those factors that affect fertility in the absence of a family planning program.

There are essentially two broad types of quasi-experimental designs that are appropriate to the evaluation of ongoing programs.

(1) Correlational designs based on cross-site program variation. Our nested forms of government provide a useful source of variation in program delivery. Thus, we can anticipate that in some states and in some local communities a program will have excellent coverage and in other places be so slight as to be almost nonexistent (and sometimes, in

fact, nonexistent because of the failure of states and local governments to opt for the program). At the present time our public welfare system is hardly uniform across states and sometimes within states. Some public welfare programs are extremely generous (e.g., New York and Massachusetts) and others are penurious beyond belief (e.g., Alabama and Mississippi). Coverage may vary from state to state, with the more generous states in this respect covering not only families whose heads are unemployable but also heads that are employable. In some places, efforts are made to publicize the welfare program eligibility requirements in order to obtain as large a coverage as possible of the eligible population. In other states, public welfare eligibility requirements are held almost as state secrets.

This variation from place to place in the intensity and coverage of treatments provides a means for evaluating effectiveness. Simply put, treatment levels that are heavy and broadcast widely among eligible populations should produce more effects than do treatment levels with the opposite characteristics, ceteris paribus. Thus, Cutright and Jaffe (1977), in their analysis of the effectiveness of the family planning program, essentially examined the fertility rates of groups of counties that had programs with wide coverage with the fertility rates of groups of counties with opposite program characteristics, holding constant county characteristics known to be related to fertility (e.g., age composition, socioeconomic level, and the like).

The phrase ceteris paribus is, of course, the obstacle to be overcome. Hence the stress on a priori understanding of the phenomenon in question.[11] It is previous knowledge about what cause inter-area variation in fertility that made it possible for Cutright and Jaffe (1977) to make other things equal statistically. It is the questioning of that knowledge in the Coleman report that produced the controversies surrounding its interpretation, with the economists claiming that Coleman had misspecified his model of how individuals varied naturally in their educational achievement levels.

(2) Time series designs based on variation over time. The second approach to the evaluation of ongoing programs rests on the existence of variations over time in the extent and intensity of treatments. Thus, changes in the level of treatment obviously occur at the start of a program, the change going from zero to an initial delivery level, and subsequent changes in policy produce variations in the amount and coverage of treatments over time.

A change in one or both respects should produce a change in a desired effect, ceteris paribus, if the treatment is effective. In a time series analysis of the effect of the 1974 Massachusetts gun control law, Deutscher and Alt (1977) found that crimes in which firearms were used declined significantly after the gun control law went into effect. His analysis took into account the long-range trends in such crimes in Massachusetts by constructing a model that fit such trends and extrapolating that model to cover the period after the gun control law went into effect. Similar analyses have been made of the effect of changes in our national labor relations laws on the incidence of strikes, and of the changes in speed limits on traffic deaths.

The ability to undertake time series analyses depends, obviously, on the existence of accurate measurements of intended program effects taken over a relatively long period of time. Thus deaths from traffic accidents, reports of crimes committed and known to the police, the incidence of industrial strikes, and fertility measures are all examples of measures for which relatively long and reliable time series exist. For other types of intended effects for which time series are not available, longitudinal analyses cannot be taken.

The technical issues surrounding the use of cross-sectional and time series designs have been dealt with at length in other publications (Hibbs, 1976). One need only summarize in the context of this paper: a variety of statistical models are available that, when appropriately employed in connection with valid substantive models, can produce firm evaluations of ongoing programs.

CONCLUSIONS

This article has attempted to provide a generalized characterization of human services and a detailed account of some of the difficulties in evaluating the effectiveness of such services. We pointed out that the critical feature of human services is that they are highly operator-dependent and difficult to standardize. Hence, it is alway problematic whether a treatment is being delivered as designed, whether the mode of delivery is adding some unintended treatment to the basic one, and finally whether a treatment can be delivered in a reasonable way at all by the typical human services organizations.

The paper also sets forth a strategy for the evaluation of human services treatments. A progressive series of tests are suggested starting

with a tight experimental design for the evaluation of the effectiveness of a treatment under the best possible mode of delivery, through a final evaluation by means of correlational designs that test the effectiveness of a human services program that has been enacted into social policy.

Although the article stresses the boobytraps and pitfalls that lie in the way of someone wishing to do evaluations of human services treatments, it is not intended to advocate an avoidance of this area of social science research. Rather, by pointing out some of the difficulties, it is hoped that the article has presented a challenge to some of the more ingenious research designers to try their hand at this rather difficult game.

NOTES

1. Policy analysis may be viewed as the application of social science theory along with the results of social science research to the examination of policy alternatives. Properly undertaken, policy analyses assume that alternative policies have been evaluated and that their effectiveness values are known.

2. This duality is closely related to that involved in the provision of public goods. It makes little sense for any individual to pay taxes unless paying taxes is made compulsory for all since the marginal utility for any individual of the majority of public services is very small.

3. The author cannot cite a public reference for this example, since knowledge of it stems from a consulting relationship with the organization that contracted to deliver the services.

4. These generalizations stem from observations made of caseworker/client transactions in four Massachusetts public welfare offices during the summer of 1976.

5. The R&D centers funded by the Office of Education, although not very successful as a group either in research or in development, were in principle a step in the right direction. They were intended as centers in which effective educational treatments would be developed and then tested out in cooperation with school systems until an effective diffusable system could be worked out. Some of the reasons for the poor performance of the R&D centers are given in Rossi (1976).

6. There are two types of contradiction possible: first, some goals are logically incompatible in the sense that the achievement of one goal makes it logically impossible to achieve another goal (e.g., reducing payments to unemployed persons, ceteris paribus, and maintaining the purchasing power of the unemployed). Second, some goals are empirically incompatible in the sense that achieving one goal empirically implies diminishing the ability to achieve another. The example given in the text is that of empirical incompatibility.

7. If the treatment is frontal lobotomy, the treatment is likely to be successful in the sense that it cures delinquent behavior, but also "cures" (or eliminates) other types of behavior as well, including much of the behavior forms that allow the individual to

function normally in the society. A treatment that is effective must not only eliminate the condition in question but also not impose other deficiencies upon the individual. Untoward side effects must also be avoided, a point that may often be overlooked.

8. The question whether an experiment of this sort tested a sufficiently wide range of the treatment is one that can always be raised post facto when results are found that indicate that treatments had no discernible effect. The question is whether the advocates of the treatment and its potential users agree a priori that the range of treatment covers what they consider to be a reasonable test of the treatments' effectiveness. As a matter of strategy, I suggest that such treatments exceed the range of reasonable treatment levels, anticipating the criticism that the trial of the treatment was unfair in the range of treatments tested.

9. It should be borne in mind that one of the motivations of the Department of Housing and Urban Development in funding the administrative "experiment" was to buy political support and time while two very well designed randomized experiments were being run: a demand experiment is testing the effectiveness of the treatments in bringing about an increase in the quality of housing occupied by families under allowance payment plans. A supply experiment would test the responses of local housing markets to the existence of payment plans, hopefully by increasing the supply of acceptable low-cost housing, as opposed to raising prices on existing housing.

10. Two outstanding examples of the use of existing records in extremely creative ways ought to be cited here: first, Cutright and Jaffe (1977) used counts of clients served in family planning clinics in groups of counties throughout the country to evaluate clinic effectiveness by relating such counts to subsequent birthrates. Second, the Vera Institute traced a large sample of felony arrests through to the final dispositions of each case, providing excellent accounts of the circumstances under which plea bargaining is used and the kinds of cases that are brought finally to trial. The Vera Institute researchers also conducted intensive studies of subsamples of cases, interviewing the states' attorneys and defense attorneys in these cases in order to obtain an understanding of the decision processes used.

11. The issues involved in specification errors have been thoroughly reviewed in Cain (1975).

REFERENCES

BLAU, P. M. (1955) Dynamics of Bureaucracy. Chicago: Univ. of Chicago Press.

CAIN, G. G. (1975) "Regression and selection models to improve non-experimental comparisons," pp. 297-317 in C.A. Bennett and A.A. Lumsdaine (eds.) Evaluation and Experiment. New York: Academic.

CARLSON, D. B. and J. D. HEINBERG (1977) How Housing Allowances Work. Washington, DC: Urban Institute.

CLINE, M. G. (1975) Education as Experimentation: Evaluation of the Follow Through Planned Variation Model. Cambridge: Abt Associates.

CUTRIGHT, P. and F. S. JAFFE (1977) Impact of Family Planning Programs on Fertility: The U.S. Experience. New York: Praeger.

DAVIS, J. A. (1962) Stipends and Spouses. Chicago: Univ. of Chicago Press.
DEUTSCHER, J. and F. B. Alt (1977) "The effect of Massachusetts' gun control law on gun-related crimes in the city of Boston." Evaluation Q. 1: (November): 543-567.
GRAMLICH, E. M. and P. P. KOSHEL (1975) Educational Performance Contracting. Washington, DC: Brookings Institution.
HIBBS, D. A., Jr. (1976) "On analyzing the effects of policy interventions: Box-Jenkins and Box-Tiao vs. structural equation models," in D. L. Heise (ed.) Sociological Methodology, 1975. San Francisco: Jossey-Bass.
KASSEBAUM, G., D. WARD, and D. WILNER (1971) Prison Treatment and Parole Survival. New York: John Wiley.
McLAUGHLIN, M. W. (1975) Evaluation and Reform: The Elementary and Secondary Education Act of 1965/Title I. Cambridge: Ballinger.
REISS, A. J. (1971) The Police and the Public. New Haven: Yale Univ. Press.
ROSSI, P. H. (1976) "Assessing organizational capacity for educational R&D in academic institutions." Educational Researcher 5.
——— and K. LYALL (1976) Reforming Public Welfare. New York: Russell Sage Foundation.
ROSSI, P. H. and S. R. WRIGHT (1977) "Evaluation research: an assessment of theory, practice and politics." Evaluation Q. 1: 5-52.
ROSSI, P. H., R. A. BERK, and B. K. EIDSON (1974) The Roots of Urban Discontent. New York: Wiley-Interscience.
U.S. Department of Labor (1977) Unlocking the Second Gate: R&D Monograph No. 45. Washington, DC: Government Printing Office.

Peter H. Rossi is Professor of Sociology and Director of the Social and Demographic Research Institute at the University of Massachusetts, Amherst. He is Editor of Social Science Research and President-Elect of the Evaluation Research Society. He is also coauthor (with Walter Williams) of Evaluating Social Programs *(Academic Press, 1972).*