University of Wisconsin-Madison
Institute for Research on Poverty

Focus

Volume 12

Number 4

Fall 1990

ISSN: 0195-5705

Where we are in the evaluation of federal social welfare programs

by Charles F. Manski
Director, Institute for Research on Poverty

It may seem self-evident that social welfare programs should regularly be assessed and refined in the light of lessons drawn from experience. Nevertheless, systematic program evaluation is a recent development. Modern evaluation practice is generally agreed to have begun in the middle 1960s, when initial attempts were made to evaluate programs enacted or proposed as part of the War on Poverty.[1] Evaluation has since spread rapidly; today, almost every substantial social program is subjected to some form of evaluation.[2] At the same time, evaluation has evolved into both a professional discipline and an industry.[3]

There now exists a consensus that program evaluation is important and should be an integral part of the policy process. But there is no consensus on the manner in which evaluations should be performed and the way their findings should be interpreted. At the moment, the most heated controversy concerns the relative merits of statistical analysis of controlled social experiments and econometric analysis of actual program outcomes. A less visible, but simmering, dispute questions the logic of the traditional distinction between "process" and "impact" evaluations.

Enactment of the Family Support Act of 1988 makes it timely to ask where we are in the evaluation of federal social welfare programs. The Family Support Act will be the focus of evaluation efforts in the next several years. Under Title II, the Job Opportunities and Basic Skills Training program (JOBS), Congress mandated separate "implementation" and "effectiveness" studies of training programs initiated by the states under the Act.[4] Taking unusually specific action,

Congress even stipulated the mode of data collection for the effectiveness study: "a demonstration project conducted under this subparagraph shall use experimental and control groups that are composed of a random sample of participants in the program."[5]

Concern with the Family Support Act in particular and with program evaluation in general led the Institute for Research on Poverty and the Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services, to jointly organize a major conference, "Evaluation Design for Welfare and Training Programs," held in April 1990. The proceedings of the conference are summarized later in this issue of *Focus*. The papers commissioned for and delivered at the conference will be published in a forthcoming book, *Evaluating Welfare and Training Programs*, edited by Charles F. Manski and Irwin Garfinkel (Harvard University Press, 1991).

The present article offers one person's perspective on the evaluation of federal social welfare programs. To focus the discussion, I first present a flowchart describing an important class of federal programs. I then describe and critically assess current practice in evaluating such programs. The article concludes with recommendations for improving evaluation practice.

Schematic of a federal social welfare program

Figure 1 outlines a typical federal social welfare program. Three existing programs, Aid to Families with Dependent Children (AFDC), the Job Training Partnership Act (JTPA), and Unemployment Insurance (UI), share this structure.

Arrows 1, 2, and 3b trace the process by which a program is fleshed out. Federal statutes and regulations sketch the program, leaving a state with substantial discretion in the way it will comply with the federal mandate. Negotiations between the state and the federal government yield an accepted state program. The state-federal agreement specifies major program provisions but inevitably leaves many details to be tied down by the state as it administers the program. Program administration may itself be a multi-tiered process, involving state, county, and local agencies as well as private service providers; this subprocess is omitted from the figure for the sake of simplicity.[6] In the end, decisions about program eligibility and specification of treatments may be made by individual caseworkers in local welfare offices and by service providers operating under government contract.

Arrows 3a, 3b, and 3c describe the determination of program participation. A participant emerges from the population when a potentially eligible person applies for entry into the program. Eligibility is not determined solely by the program's formal rules; in practice, the rules are interpreted by local officials. Moreover, in many cases, initially ineligible persons may become eligible by modifying their behavior appropriately.[7] The participation process takes place in an environment shaped by the local economy and social norms. In particular, a person's economic options and the social stigma associated with program participation will be influential as a person decides whether to become eligible and to apply to the program.

Arrows 4, 5a, 5b, and 5c show the program's potential impacts. The term "impact" is sometimes applied only to the program's direct effect on participants. But attention must also be given to feedback effects, shown in the dotted arrows. For example, a training program may have effects on the operation of the labor market; a program for the homeless may affect the housing market. The social stigma associated with a program may change with the number of persons who participate. A state may revise the way it administers a program as it observes how the program affects participants. A program may even alter the composition of the population of an area; for example, it is often asserted that a relatively liberal AFDC program makes a state a "welfare magnet."

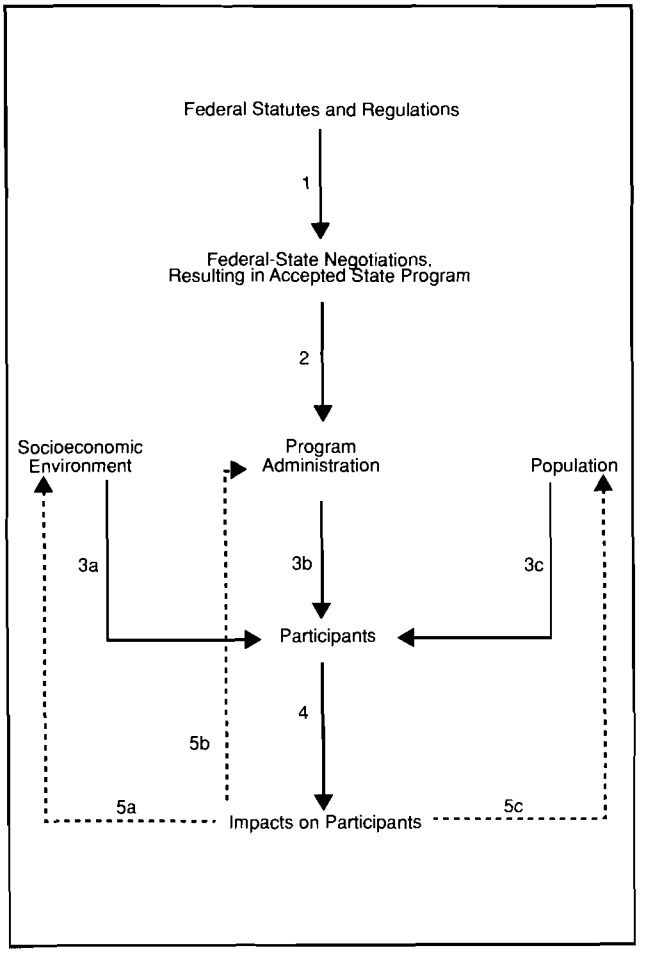

Figure 1: A Federal Social Welfare Program

Note: Dotted lines indicate feedback effects.

Evaluation practice

In principle, an evaluation of a federal social welfare program should seek to illuminate the entire complex process depicted in Figure 1. To be useful in policy formation, an evaluation should seek to answer *counterfactual* questions: what would change if some aspect of a program were altered? In practice, evaluators inevitably simplify this daunting task. Three major simplifying features of current evaluation practice follow.

- Restriction of the domain of counterfactual analysis

Of all the processes shown in Figure 1, the only ones regularly subjected to counterfactual analysis are program participation (arrows 3a, 3b, and 3c) and the direct impact of a program on participants (arrow 4). The federal-state negotiation of an agreed program (arrow 1) is generally ignored entirely.[8] Program administration (arrows 2 and 5b) is subjected only to process evaluation, traditionally a descriptive exercise rather than a systematic comparison of the existing program administration with alternatives. Feedbacks to the socioeconomic environment and to the population (arrows 5a and 5c) are sometimes noted as possibilities, but are almost always ignored in actual evaluations.

A striking feature of evaluation practice is its disciplinary specialization. Process evaluation of program administration is the province of qualitatively trained political scientists. Program participation and impacts are analyzed primarily by economists, almost always using some quantitative approach.

- Separation of impact analysis from participation analysis

Many analyses of program participation and impacts seek to interpret "natural variation" across programs: either cross-sectional variation in outcomes across states with different versions of the program or time-series variation within a state that alters its program. It is by now widely recognized that the analysis of natural variation requires the evaluator to analyze program participation and impact jointly. The people who choose to become eligible and apply to a program are presumably those who expect the program to have a favorable impact on them. Provided only that expected impacts are related to actual ones, program participation and impacts are jointly determined outcomes.[9]

The use of natural-variation data to jointly analyze program participation and impact is generally agreed to pose a difficult scientific task.[10] As a consequence, evaluators who wish to analyze impacts and are not concerned with participation per se have often turned to "controlled social experimentation" as a mode of data collection. In the typical controlled social experiment, persons who apply to a program are randomly assigned to different versions of the program, perhaps including "non-treatment." Random assignment ostensibly breaks the tie between participation and impact that is inherent in natural-variation data. Hence, the evaluator can study impacts without having to jointly analyze participation. Thus, controlled social experimentation promises a substantial simplification of the evaluation problem.

- Industrialization and standardization of program evaluation

The major program evaluations of the 1960s and early 1970s were designed and performed by academic researchers, in collaboration with early evaluation professionals. While small-scale analysis continues to take place in universities, large-scale program evaluation has increasingly become the domain of private firms specializing in such endeavors.

The emergence of an evaluation industry has been accompanied by increased standardization in the design, performance, and presentation of findings from evaluations. Standardization is most notable in the analysis of direct program impacts, as the evaluators of social welfare programs have sought to emulate the routinized controlled-experimentation procedures of the physical and biological sciences.

Weaknesses in current practice

The inherent complexity of program evaluation makes efforts at simplification essential. At the same time, we must be careful not to simplify away essential aspects of the evaluation task. I believe that current evaluation practice sacrifices too much in the name of simplification. Several weaknesses now limit the usefulness of our evaluations.

Failure to recognize that process is part of treatment

The distinction between process and impact evaluation, albeit long-standing, is untenable. A federal social welfare program is not a complete set of procedures whose implementation can be monitored and controlled perfectly. In reality, a federal "mandate" to the states only establishes a set of rules and incentives intended to influence the behavior of the states. Similarly, a state cannot perfectly monitor and control the administration of a program; it can only establish a set of rules and incentives intended to influence the behavior of the local agencies and service providers that ultimately carry out the program. The lesson is that, from the perspective of federal policymaking, a program is not defined solely by its treatment of participants; it is defined as well by its treatment of state governments, local agencies, and service providers. Hence process is part of treatment.

The established practice of separating process and impact evaluation has adverse consequences. Process evaluations only describe program administration, but policy formation requires answers to counterfactual questions. We need to know how program outcomes would change if the rules and incentives given to states, local agencies, and service providers were altered. For example, how would states change

their existing programs under JTPA if the federal government were to alter the performance standards now in place?[11] How will states change their JOBS programs when, as expected, performance standards for this program are eventually enacted?[12] How do the job training and basic education services provided by private contractors under JTPA and JOBS change as a function of the prevailing payment formula?

The failure to recognize that process is part of treatment also has troubling implications for the interpretation of findings from controlled social experiments. Findings from an experiment are useful only if program administration under the experiment does not systematically differ from administration in a full implementation of the program. There are many reasons to question this premise. Small-scale experiments typically do not produce the same local caseloads as do full program implementations. Nor do they provide caseworkers and program participants with the same information about program features and impacts. Particularly problematic is the fact that social experiments cannot be performed using the double-blind protocols of medical trials, in which neither experimenter nor subjects know who is in each treatment group. Caseworkers and service providers necessarily know who is in each treatment group and cannot be prevented from using this information to influence outcomes.

Inappropriate extrapolation from controlled social experiments

The assumption that the program administration observed in a controlled experiment will remain unchanged when the program is implemented fully is one of several common but inappropriate extrapolations from experiments to the real world. Another is the widespread assumption that the pool of applicants to an experimental version of a program will remain unchanged when the full-scale version of the program is implemented. This is not plausible, because the private value of applying to a program with randomized treatments is not the same as that of applying to a program with known treatment.[13] A third improper extrapolation arises from the practice in social experiments of ignoring feedbacks from the program to the socioeconomic environment and population. The scale of the typical social experiment may be too small to discern feedback effects that become prominent when the program is implemented fully.[14]

The difficulty of extrapolating from an experiment to the real world has long been known. Extrapolation problems arising in the social experiments of the 1970s led evaluation researchers of that period to become cautious in interpreting experimental evidence.[15] Unfortunately, the lessons of the 1970s seem not to have been learned by today's social experimenters. Among this group, a proper awareness of the difficulty of natural-variation analysis has often been accompanied by an overly sanguine view of experimentation. Some have gone so far as to assert that only experimental evidence should be used to evaluate social welfare programs.[16] It is important to recognize that deep problems hinder the interpretation of both experimental and natural-variation data.

Lack of balance between applications and basic research

From the mid-1960s through the late 1970s, evaluations of social welfare programs nicely blended applications and basic research. Specific programs were analyzed and policy implications drawn. At the same time, innovation in evaluation methods took place and a base of empirical knowledge guiding future evaluations was established. Social scientists, evaluation professionals, and public officials not only worked together but sometimes traded hats.

In the past decade, funding for basic evaluation research has substantially diminished. Simultaneously, the public has increasingly demanded proof of the effectiveness of existing and proposed social programs. The consequence is that evaluation today is dominated by tightly focused applications with short horizons. Government and foundation funding is allocated largely through contracts calling on the evaluator to provide specified deliverables on a fixed schedule. The contractor's task is usually to compare the short-run direct impact of a given program with that of a particular alternative.

Restoration of the balance between applications and basic research is sorely needed. The existing environment has clear negative implications for the long-term health of evaluation practice. Present contractual funding promotes unimaginative evaluations, executed using conventional procedures, reported in a standardized format. It discourages innovation in methods, stifles efforts to understand the complex set of processes that define a program, prevents evaluation of long-term program impacts, and inhibits creative thinking about the design of new programs.

Recommendations

These weaknesses in current evaluation practice indicate the need for changes:

1. The conventional separation of process and impact evaluation should end. The operational definition of program treatment should be expanded to include not only the treatment of participants but also the treatment of state governments, local agencies, and service providers. Evaluations should seek to answer counterfactual questions about all the dimensions of treatment.

2. The assertion that evidence from controlled social experiments is qualitatively superior to natural-variation data should be dismissed, as it is not supportable. Program evaluations should employ both experimental and natural-variation data, in all cases with due caution.

3. The present funding imbalance between applications and basic evaluation research should be corrected. Effective collaboration of social scientists, evaluation professionals, and public officials once made the evaluation of federal social welfare programs a creative enterprise with both immediate and long-term benefits to society. This collaboration must be renewed. ■

---

[1]See, for example, Henry Aaron, *Politics and the Professors* (Washington, D.C.: The Brookings Institution, 1978), p. 30; Robert Haveman, *Poverty Policy and Poverty Research* (Madison, Wis.: University of Wisconsin Press, 1987), chap. 8; and Robert Lampman, "The Decision to Undertake the New Jersey Experiment," Foreword to David Kershaw and Jerilyn Fair, *The New Jersey Income-Maintenance Experiment* (New York: Academic Press, 1976). Earlier evaluation efforts were largely limited to "process" evaluations, describing the administration of a program. Modern evaluation practice does, however, have some historical precursors. For example, the Children's Bureau of the U.S. Department of Labor studied the impact on infant mortality of the Sheppard-Towner Act, a 1921 statute establishing infant nutrition programs (see *Report of the Committee on Public Health Organization, Section II: Public Health Service and Administration* [New York: Century Co., 1932] and U.S. Department of Labor, Children's Bureau, *The Seven Years of the Maternity and Infancy Act* [1931]). I am grateful to Linda Gordon for bringing this early evaluation to my attention.

[2]The recent activity of the federal Interagency Low Income Opportunity Advisory Board is revealing. The Board, created by Executive Order of President Reagan on July 20, 1987, was established in part to review state proposals for welfare reform demonstrations under AFDC, Food Stamps, and other federal social welfare programs. Once constituted, the Board decided to require that every state demonstration proposal be accompanied by an evaluation plan designed to measure the net effects on dependency and the cost-effectiveness of the demonstration. Acceptance of the evaluation plan became part of the approval process for a demonstration (see Michael Fishman and Daniel Weinberg, "The Role of Evaluation in State Welfare Reform 'Waiver' Demonstrations," in *Evaluating Welfare and Training Programs*, ed. Charles F. Manski and Irwin Garfinkel, forthcoming, Harvard University Press).

[3]There now exist professional journals devoted to evaluation, including *Evaluation Review* and *Evaluation Forum*, as well as a professional society, the American Evaluation Society. Courses in evaluation are offered routinely in the public policy schools of universities throughout the country. The evaluation industry includes such large firms as Abt Associates, Lewin/ICF, the Manpower Demonstration Research Corporation, and Mathematica Policy Research, among others.

[4]The terms "implementation" and "process" are roughly synonymous, as are "effectiveness" and "impact."

[5]Public Law 100-485, October 13, 1988, Section 203, 102 Stat. 2380. A contract to perform the mandated effectiveness study has been awarded to the Manpower Demonstration Research Corporation. The Secretary of the U.S. Department of Health and Human Services has also appointed an Advisory Panel for the Evaluation of the JOBS Program, composed of public officials and academic experts.

[6]JTPA gives states and localities an especially large degree of latitude in program design. This is discussed by V. Joseph Hotz in "Recent Experience in Designing Evaluations of Social Programs: The Case of the National JTPA Study," in Manski and Garfinkel, *Evaluating Welfare*.

[7]For example, a woman can choose to become eligible for AFDC through her marriage, childbearing, and labor supply decisions. A worker can choose to become eligible for UI by not accepting an employer's offer of an out-of-state transfer following a plant closing.

[8]Fishman and Weinberg, in "The Role of Evaluation in State Welfare Reform 'Waiver' Demonstrations," provide an informative description of a set of recent federal-state negotiations.

[9]See James Heckman and Richard Robb, "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. Heckman and Burton Singer (New York: Cambridge University Press, 1985); and Robert Moffitt, "Evaluation Methods for Program Entry Effects," in Manski and Garfinkel, *Evaluating Welfare*.

[10]There is, however, considerable debate concerning the seriousness of this difficulty. For two opposing views, see Robert LaLonde, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76 (1986), 604-620; and James Heckman and V. Joseph Hotz, "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, no. 408 (1989), 862-874.

[11]Burt Barnow, "The Effects of Performance Standards on State and Local Programs: Lessons for the JOBS Program," in Manski and Garfinkel, *Evaluating Welfare*, describes in detail the existing JTPA standards and speculates on their effects on state and local behavior.

[12]Section 203 of the Family Support Act requires that the Secretary of the U.S. Department of Health and Human Services submit recommendations for performance standards to Congress by October 1, 1993.

[13]This point is developed forcefully in James Heckman, "Randomization and Social Policy Evaluation," in Manski and Garfinkel, *Evaluating Welfare*. In principle, the problem can be avoided by offering treatment to a random sample of the general population rather than to a random sample of program applicants. In practice, cost considerations have always led experimenters to randomize applicants.

[14]This problem is discussed in detail in Irwin Garfinkel, Charles F. Manski, and Charles Michalopoulos, "Micro Experiments and Macro Effects," in Manski and Garfinkel, *Evaluating Welfare*.

[15]See *Social Experimentation*, ed. Jerry Hausman and David Wise (Chicago: University of Chicago Press, 1985).

[16]See, for example, Laurie J. Bassi and Orley Ashenfelter, "The Effect of Direct Job Creation and Training Programs on Low-Skilled Workers," in *Fighting Poverty: What Works and What Doesn't*, ed. Sheldon H. Danziger and Daniel H. Weinberg (Cambridge, Mass.: Harvard University Press, 1986) and LaLonde, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." It is worth noting that similar assertions have been made in the health field. For example, a recent National Research Council study of AIDS prevention programs, citing the difficulty of interpreting natural-variation data, has asserted that only evidence from controlled experiments should be used to evaluate such programs (see *Evaluating AIDS Prevention Programs*, ed. Susan L. Coyle, Robert F. Boruch, and Charles F. Turner [Washington, D.C.: National Academy Press, 1989]).