Evaluating comprehensive family service programs: Conference overview

Since 1986 a number of federal agencies have initiated large-scale demonstration programs designed to relieve the deprivation of parents and children in impoverished families. The evaluation of such programs formed the subject of a conference held in Washington on November 14–15, 1991. The conference was one of a series jointly sponsored by the Institute and the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in the U.S. Department of Health and Human Services.

This conference series began in 1989 with a one-day workshop in Washington to provide ASPE staff and other members of executive departments with expert counsel on practical approaches to evaluating the programs created by the Family Support Act. The second meeting, more academic in tone, consisted of a two-day national conference in 1990 at Airlie House, Airlie, Virginia, where federal representatives, evaluation professionals, and academic researchers examined the assessment of welfare and training programs.[1]

The 1991 conference advanced into the more complex realm of projects offering services for disadvantaged parents and their children. Represented among its 120 participants and observers were academicians, professional evaluators, federal staff involved with program planning, and members of private foundations and service organizations. The programs discussed at the conference are those sometimes referred to as "two-generation interventions":

> A potentially powerful new strategy for assisting families in poverty is being tested in a set of new program models that target welfare-dependent women with young children. These models vary in several respects, but have a common strategy: they help families attain economic self-sufficiency through education and job training while also providing other services, such as parenting education and high-quality child care, that support children's healthy development. As two-generation interventions these programs show promise of addressing both immediate and long-term impediments to healthy development and educational success for poor children.[2]

By attempting to improve simultaneously the circumstances of parents and the life chances of their children, these programs span welfare and employment efforts on one hand and child development, child welfare, and social service efforts on the other, with the result that those operating the programs as well as those evaluating them represent a variety of disciplines and professions.

The evaluation projects for seven major programs were presented and discussed. Three of the programs were authorized by Congress (the Job Opportunities and Basic Skills Training Program, JOBS; the Comprehensive Child Development Program, CCDP; and the Even Start Family Literacy Program). Two originated in federal executive agencies (the Teenage Parent Demonstration and Youth Opportunity Unlimited, YOU). One is a state initiative (the Washington State Family Independence Program, FIP), and one (New Chance) is privately sponsored. In addition, programs still in early stages were briefly discussed, and the evolution of Head Start evaluations over the past twenty-five years formed the subject of a special presentation. Capsule descriptions of the programs and their evaluations accompany this article, and main features of the seven large projects are compared in Table 1, pages 14 and 15. The conference agenda appears on page 21.

The conference had four principal purposes: to summarize the state of evaluation methodology, to identify the key issues in assessing these complicated programs, to permit evaluators in different fields and disciplines to pool their knowledge, and to help ASPE structure future evaluations in the area of family services. The consensus, upon conclusion, seemed to be that while the meeting moved forward on all four dimensions, a significant contribution lay in providing the opportunity for evaluation contractors to exchange information concerning the nature, problems, and accomplishments of their projects.[3] Also important was the opportunity for federal staff members from the legislative branch (Senate and House, General Accounting Office, Congressional Budget Office) as well as the executive branch (Housing and Urban Development, Education, Labor, Agriculture, several agencies within Health and Human Services, the Census Bureau, and the Office of Management and Budget) to attend the deliberations and gain knowledge bearing on their own work.

Several themes materialized during the presentations and the vigorous discussions that ensued. The following summary attempts to capture major points.

The time and place for evaluations

Martin Gerry, Assistant Secretary for Planning and Evaluation in DHHS, noted in his introductory remarks that formal evaluation of social programs has taken on greater importance in recent years, amid growing concern over learning what works, and how well. Evaluation in the 1980s

of experiments by several states with welfare reform directly influenced the Family Support Act and encouraged Congress to include evaluation requirements in the authorizing legislation for two other programs (see capsule descriptions).

Conference participants noted the advantages that can accrue from a congressional mandate for evaluation. It strengthens the hand of government researchers who want to analyze the effects of public policy on individual behavior. It may open doors to funding by government agencies that would otherwise remain closed. And the specification of a particular form of evaluation, exemplified in the requirement that random assignment be used for the JOBS evaluation, can help researchers convince reluctant program operators that there is reason to assign clients to different forms of treatment.

The problem with this congressional attention, pointed out in other comments, is that it may impede evaluation design. The federal procurement process that is set in motion by a congressional mandate for evaluation sometimes occurs too early, before a program has been clearly developed—before there is certainty concerning what is to be evaluated. Allowance must therefore be made for changing the evaluation design as the program matures and alters. This may be accomplished by explicitly permitting and encouraging redesign as a program progresses.

In the case of programs whose effectiveness is contested and controversial, as is true of those involving family preservation services, it may be desirable first to step back and assess the feasibility of an evaluation before proceeding to design one. In other cases, evaluation can profit from prior experience and move to a second generation of effects, comparing not just the average effect of Treatment A among all those who receive it versus those who do not, but the relative effects on different subgroups of Treatment A versus Treatment B. In this way the JOBS evaluation benefited from the years of experience that preceded it, when the Manpower Demonstration Research Corporation (with support from private, not public, funds) evaluated state experiments in welfare reform.

Such experience is sorely lacking in other program areas, especially in the complicated realm of family services. Conference members agreed that careful thought is required in advance to identify the subject of evaluation, the variables to be defined, and the measures to use. And yet, as one participant commented, too much delay in formulating an evaluation may mean that it never gets off the ground.

The design of the evaluations

The opening remarks that described the charge of the conference called attention to the fact that the seven major evaluation projects share several design features. All but one (YOU) use random-assignment designs to measure effects on parents. Of these all but one (FIP) measure effects on children as well. In an effort to determine what dimensions of a program make a difference—what works for whom—increasingly complex experimental designs are being used, such as the random assignment scheme of JOBS. For similar reasons, most of the evaluations are collecting a large amount of baseline information prior to random assignment. This information often goes beyond simple demographic variables to include, as do New Chance and JOBS, measures of depression, baseline literacy, and self-confidence.

All of the evaluation designs include cost-benefit analyses. This is an especially difficult exercise when programs provide benefits that are hard to quantify. How can one give a monetary value to benefits that children obtain from the education and training of their parents?

Another universal feature of these evaluations is that they contain extensive studies of implementation: that is, they closely observe what services are delivered to which clients and how the service delivery system is organized. This scrutiny of what goes on "inside the black box" to learn about the intensity of services, the structure of services, and the details of staff-client interactions should reveal not only how programs shape people, but how people shape programs.

On the other hand, certain design characteristics are unique to individual projects. FIP matches treatment and comparison sites, rather than randomly assigning clients within sites. YOU allocates funds to neighborhoods rather than to service projects. CCDP assigns special observers to record program implementation at each site. Even Start focuses on adult and child literacy and their interconnection. Three projects—the Teenage Parent Demonstration, JOBS, and New Chance—have embedded more detailed, qualitative substudies within the larger evaluation. Some of the projects require clients to participate; others are voluntary.

Particular design issues that were discussed extensively at the conference include (1) designating the unit of analysis, (2) determining the appropriate follow-up period, and (3) taking account of "transactional analysis," defined below.

What is the focus of analysis?

The problem of defining the unit of analysis is endemic in these two-generation programs, owing to the many actors involved. The teen-parent intervention directly concerned mothers and their children, but also affected the lives of others—parents of the mother, other relatives, boyfriends, the children's fathers. The question becomes which units to track in the course of evaluation. In some of the other studies, a "focus" child within a family is chosen for in-depth examination. But could we not gain rich information by examining siblings as well? Other family members? The questions remain open.

How long should an evaluation last?

Determining an appropriate follow-up period is also difficult. The two-year follow-up for the teen-parent impact analysis means that the average age of sample members will then be 20, yet the transition from school to work usually covers ages 18 to 24. Would it not be preferable to extend the follow-up to a longer span of time? This is an expensive proposition, and adequate resources may not be available. In the case of programs such as Even Start that involve early childhood education, we would like to know what happens to the children as they progress through elementary and secondary school. In the case of programs intended to improve parenting skills, like CCDP and New Chance, we want to learn what kinds of parents the children themselves become, one generation later. The time horizon stretches on.

Can we learn more about behavioral changes?

Reference to the interaction of case managers and their teen-parent clients prompted a recommendation from psychologists at the conference that evaluation of these programs should give consideration to transactional analysis, a term referring to study of the succession of modifying interactions that take place in the course of a program—between managers and clients, between mothers and children, among the various agents involved in the process of a program. This form of analysis is dynamic, going beyond observation of single individuals at fixed points in time. Economists in the audience noted the parallels between this type of study and that described in the job-search literature, which focuses on the sequential decisions made by job seekers who solicit and receive a series of job offers. In the same way, transactional analysis follows a conditional-probability strategy—examining a particular event in the light of events that preceded it—to track the quality and cumulation of program effects.

Qualitative and observational research

Common among these evaluations is the specification of an ethnographic or observational component, a topic that received particular attention at the conference. A special study within the Teenage Parent Demonstration, funded by private foundations and about to be fielded, will examine parent-child interactions to determine the effects of the demonstration on parenting skills and child development. Its data include videotapes, interviews, and surveys of home environment. The JOBS evaluation contains a substudy of a group of mothers and children to examine family environment and dynamics. It also proposes to videotape mother-child interaction. For its process evaluation CCDP assigns to each site "project ethnographers" charged with providing descriptions of the dynamics and natural history of the unfolding projects. Even Start measures a parent's ability to teach a child by observing a particular "task": while the parent reads a simple book to the child, a trained observer uses a precoded rating form to record aspects of their interactions. YOU calls for periodic, intensive field visits by trained ethnographers to describe the nature of community life, problems encountered in delivering services, and the experiences of youth in the program.

The value of this kind of information was underscored by conference participants. It offers us a closer look into the black box of program implementation, providing another layer of explanation concerning a program's operation and effects. It illuminates differentials in treatments, helping us discern when a program is well managed or when its clients are ill served. It permits appreciation of the richness and complexities of the experiences of staff and clients in these multifaceted programs. Not least important, it offers accessible, even colorful, information to the program sponsors, members of government at all levels, and the concerned public. This type of data sustains interest in a project until outcome data are available, which often takes three to five years.

Some critics took issue with this form of research. Terminology was one target: "ethnography" in its strictest sense refers to a branch of anthropology dealing with systematic description of human cultures according to prescribed procedures. This is not necessarily the sense in which the term is applied in the evaluations, even though it appears in their descriptions. "Observational research" may be a more accurate term, but its results are just that—observations made by individuals, potentially carrying a subjective element, no matter what pains are taken to reduce that element through careful training of the observers and use of standard protocols for observation.

The utilization of qualitative data is fraught with difficulties. Many of the research contractors and government project officers acknowledged that they face a formidable task as they attempt to merge process data with outcome data to gain understanding of what works for whom, and why. General agreement prevailed that information of this nature has value and purpose but must be collected and used with care and precision.

The need to extend basic research and disseminate its results

Prominent themes in the discussions included the need for standardization of measures, for "meta-analyses," and for syntheses of research results.

The multiplicity of units and variables factored into these evaluations means that further research should be devoted to ways in which to measure effects and to specify which effects we want to measure. Standardization of measures is a basic requirement if we are to draw generalizable conclusions from these assessments. There is little uniformity across programs, for example, on measurement of program participation. Is it a specified percentage of time spent in

program activities over a specified calendar period? Should it include a measure of intensity of participation? How does one gauge intensity? A large challenge to the JOBS evaluation lies in formulating measures of participation that will permit comparability across sites in order to meet the required performance standards.

A consensus emerged that a coherent set of common baseline and outcome measures, of process and participation measures, would be of immense benefit. More particularly, it was recommended that analysts attempt to designate "marker variables"—basic definitions and measures common to diverse programs—to help move evaluation methodology forward by permitting convergence of analytic concepts and tools.

Meta-analysis has been defined as "the use of formal statistical techniques to sum up a body of separate (but similar) experiments."[4] As a scientific tool it has proved controversial. Its advocates argue that it can illuminate the nuggets of truth lying under a mountain of sometimes conflicting research results. Its detractors rejoin that only under severely restricted conditions can such analysis be performed well enough to be convincing. If it is indeed possible to succeed with this form of study, these complex programs offer unusually fertile ground for its application.

Several participants emphasized the need to synthesize and disseminate the results of evaluations of previous programs before launching major new efforts. An example cited was the publication of *From Welfare to Work*, a summary of the results of state experimentation with welfare reform prepared by the Manpower Demonstration Research Corporation, which provided the basis for the JOBS evaluation. (Preparation of the summary was required under the JOBS evaluation contract, as a result of a recommendation at the 1989 IRP/ASPE workshop, mentioned earlier.)

Summaries of this nature would promote dissemination of findings and provide the opportunity for evaluators to take time to think about basic issues before moving ahead. Evaluators expressed the desire for government agencies such as ASPE to commission more syntheses destined for two separate audiences: the policy community, including federal, state, and local government staff, legislative staff, advocacy groups; and the academic community. The first audience needs summaries of results for its immediate purposes. The second can use them to promote accumulation of a body of knowledge and to further the development of social science. Needed for this purpose also, it was felt, are public use tapes from the evaluations, which will facilitate secondary analyses and additional academic research.

Afterword

The conference offered testimony both to the advances that have been made in evaluating antipoverty programs and to the distances that remain to be crossed. The personal views presented below (see pp. 22–34) provide more detail concerning these achievements and deficiencies. The reflections of the three members of the academic community point to the remarkable degree of technical competence revealed by the evaluations and to the pressing need to bring to them more basic knowledge and research. The comments of members of the policy community tell us of the practical problems inherent in assessments of this nature and possible means to deal successfully with those problems.

It is hardly surprising that evaluations of two-generation interventions contain shortcomings, in view of the scope and complexity of these programs. What might be considered surprising, however, was the strength of personal concern and professional commitment expressed by virtually all conference participants. Evaluators and project officers alike repeatedly gave evidence of their solicitude for, and determination to alleviate, the circumstances of troubled families. Given that level of commitment, as well as the intellectual resources apparent in the conference room, one might conclude that we have grounds for optimism concerning efforts to overcome barriers to evaluation of complex social programs. ■

---

[1]A selected set of the papers was subsequently edited and published as *Evaluating Welfare and Training Programs*, ed. Charles F. Manski and Irwin Garfinkel (Cambridge, Mass.: Harvard University Press, 1992). See display box, page 36.

[2]Sheila Smith, "Two-Generation Program Models: A New Intervention Strategy," *Social Policy Report* (of the Society for Research in Child Development), 5:1 (Spring 1991), p. 1.

[3]Evaluation contractors are the private firms that conduct evaluations under contract with government agencies and private foundations. For a discussion of their role, see the Introduction to *Evaluating Welfare and Training Programs*.

[4]Charles Mann, "Meta-Analysis in the Breach," *Science*, Vol. 249, August 3, 1990, p. 476.

Table 1

Characteristics of Family Service Programs Being Evaluated

	JOBS	Comprehensive Child Development Program (CCDP)	Even Start	Teenage Parent Demonstration	Youth Opportunities Unlimited (YOU)	Washington State Family Independence Program (FIP)	New Chance
Status	Ongoing program	Demonstration	Demonstration	Demonstration	Demonstration	Demonstration	Demonstration
Coverage	Broad, affecting large segment of the welfare caseload (but with specially targeted subgroups)	Broad: family must have income lower than poverty level and a newborn infant or pregnant woman	Selective: family must have an adult eligible for Adult Basic Education, a child between ages 0–8, and live in a Chapter 1 attendance area	Broad, focusing on teenage custodial parents with only one child (or pregnant with first child)	Broad: affecting all youth within designated target areas of 25,000 population	Entire public assistance caseload, alternative to AFDC, in 5 sites	Selective within a highly targeted segment of the welfare caseload (parents aged 16–22 who are dropouts and gave birth by 20)
Participation Requirement	Mostly mandatory; likely to be substantial variation across sites	Voluntary	Voluntary	Mandatory; noncompliance results in a sanction that is lifted only when teen comes back into compliance	Voluntary	All families eligible for ADFC must enter FIP instead. All may then participate in employment and training (E&T)	Voluntary in most locations
Level of Disadvantage of Participants	Mixed: some are short-term recipients; others are highly disadvantaged	High: poor	High: low-literate and poor; 78% high school dropouts, 71% incomes under $10,000	High: all are teenage welfare recipients in inner-city areas. Even though one-third had completed high school and another one-third were in school, basic skills levels were very low	Mixed: depending upon community, which must have at least 30% of population below poverty	All are recipients of public assistance (AFDC-eligible)	High: nearly all young mothers without diplomas who are dropouts
Participation Rates	Modest levels of participation anticipated due to normal welfare dynamics and limited state resources for services and follow-up	Participation is voluntary; expected to vary across sites	Participation is voluntary; expected to vary across sites	Fairly high levels of at least initial compliance, but also fairly high levels of sanctioning	Modest levels initially; design intention is to reach the needs of all youth in the target area	All participate in income assistance part of FIP. Over half of those voluntarily participate in E&T part of FIP	Fairly high levels of participation due to rich services and voluntary nature of program in most sites
Structure (agencies involved in administering the program)	Program coordinated through welfare agencies	Grantees are community-based organizations, hospitals, local education agencies, universities	Grantees are local education agencies	Program coordinated through the welfare agencies in Chicago and in Camden and Newark, N.J.	Coordinated through a lead agency (SDA or PIC); to link with a wide range of organizations and programs; operating out of a site located within the target area	Income maintenance case coordination and supportive services administered by welfare agency; E&T admin. by employment security agency	Program offered through community-based organizations, schools, and municipal organizations
Mode of Service Delivery	Mixed, with heavy emphasis on off-site education and training through referrals to existing community services	Coordination of and referral to existing services; direct provision of a mix of in-home and on-site services; extensive reliance on case workers	Coordination of and referral to existing services; direct provision of a mix of in- home and on-site services; some reliance on case workers	Mixed off-site and on-site. Referrals to existing schools, GED programs, skills training programs; all sites offered workshops and GED instruction on-site (eventually discontinued in one site due to low enrollment)	Mixed, on-site in the community-based project site; coordinated through other agencies located in the target area; some off-site	Interagency arrangements with schools, community colleges, JTPA, CBOs, etc.	Education and personal development services primarily on-site, specially designed with target population in mind; skills training primarily off-site

Uniformity across Sites (in administration, service delivery)	Low: considerable local discretion	Low	Low	Moderately high, with variation primarily in the method of delivering on-site workshops, caseload sizes, and availability of community resources	Low; considerable local discretion	High uniformity of program regulations, guidelines. Moderate variations in client interactions, priorities, E&T services	High, prescriptive model
Services	Education, skills training, work experience, job search assistance, case management, child care, transportation assistance	Health, early childhood education, employment training, life skills, case-work, parent education, literacy skills	Adult basic education, early childhood education, life skills, parenting education	Education, job search, skills training, summer employment, case management; workshops on family planning, motivation, wide range of life skills	Employment and training, education, recreation and sports, counseling, health care, social services (including drug prevention), etc.	Assessment, case coordination, special services for pregnant teens, education, job search, trng., voc. trng., OJT, parenting skills, child care, transitional child care and Medicaid. Cash incentive bonuses above welfare grant for partic. in education, training, or if employed	Education, skills training, work experience, employment preparation, career exploration/ counseling, life skills instruction, family planning and health education, personal and group counseling, pediatric and maternal health care, and parenting education
Provision of Child Care	Financial support, referrals to providers; variability in quality anticipated	Coordination with Head Start, other local preschools and day care programs, direct provision of day care or preschool services	Coordination with Head Start, other local preschools, some child care provided to enable parents to participate	Financial support, counseling, referrals to providers; on-site care at two sites; considerable variation in quality	Mostly as a supplemental service to a program	Extensive funds for child care while in FIP and for 1 year after leaving FIP owing to employment	Mostly on-site or arrangements in developmentally oriented programs
Age of Participants' Children at Intake	Usually 3 to 17, but sometimes younger	0 (prior to birth) to 12 months	0 to 8 years	0 to 3 at intake; 80% had child under 1; some participants enrolled while pregnant	Not applicable	0 (prior to birth) to age 18	0 to 5, mostly at younger end
Number of Sites	8	1989: 23 1990: 24	1989: 76 1990: 123 1991: 240	3	7	5 with FIP; 5 non-FIP	16
Evaluation Design: Random Assignment or Other	Random assignment	Experimental evaluation in all projects. Random assignment	Descriptive survey of all projects and participants; experimental evaluation in 10 purposively selected projects. Random assignment in 5 of the 10	Random assignment	Process and outcome; highly qualitative	Matched comparison sites (5 and 5)	Random assignment
Number of Subjects (experimentals and controls when random assignment)	40,000–50,000	2,500 Es, 2,500 Cs	Descriptive: 3,000 families Experimental: 1,200 Es, 1,200 Cs	6,091 (1,281 in Camden, N.J.; 1,348 in Newark; 3,462 in Chicago)	NA	Approximately 15,000 FIP, 15,000 non-FIP	2,320
Start Date and Expected End Date	October 1989–September 1997	April 1990–March 1995	January 1990–October 1993	1986–1992	July 1990–June 1995	July 1988–June 1993	January 1989–September 1995
Expected Total Evaluation Budget	$15 million[a]	$10 million	$2.9 million	$3.9 million	$1.69 million	Approximately $3 million	$12 million[b]

Source: Originally prepared by Robert Granger, MDRC, and modified for the conference.

Note: For program descriptions, see accompanying summary.

[a]Includes payments to sites to offset research-related costs.

[b]Includes site payments and site development costs.