Evaluating Welfare Reform

A Guide for Scholars and Practitioners

Douglas J. Besharov
Peter Germanis
Peter H. Rossi

The University of Maryland
School of Public Affairs
College Park, Maryland
1997

1 3 5 7 9 10 8 6 4 2

The University of Maryland
School of Public Affairs
2101 Van Munching Hall
College Park, Maryland 20742

Printed in the United States of America

Contents

Preface

The welfare bill signed by President Clinton in August 1996 could result in fundamental changes in state welfare programs. The general public and federal, state, and local officials, along with experts and advocates on the left and right, are eagerly awaiting evidence of the new law's impact (as well as that of the changes resulting from the waivers the federal government granted to states and the earlier Family Support Act-related activities).

In the coming years, a series of studies evaluating these welfare reforms will be released. Many will provide important new information about the impact of the new regime on individuals and institutions. However, many studies will also have serious flaws that will sharply limit their usefulness.

The proper use of these forthcoming evaluations requires the ability to distinguish relevant and valid findings from those that are not. This does not mean that studies must be perfect in order to be useful: Evaluation findings are only more credible or less so. Even poorly designed or executed evaluations can contain some information worth noting.

This publication outlines this emerging body of research, primarily based on evaluations of programs under the Family Support Act of 1988 and the waiver-based state welfare reform experiments. It also describes the many new evaluations being launched in response to the passage of last year's welfare reform law.

Unfortunately, most policymakers and practitioners are ill-equipped to assess this research, especially since they must

often act before the traditional scholarly process can filter out the valid from invalid findings. This is understandable, since assessing evaluation studies often requires both detailed knowledge of the programs involved and a high level of technical expertise. To help them better assess this research and glean the lessons it offers, this publication also describes the generally accepted criteria for judging evaluations.

Program "Theory": Does the program or evaluation make sense in light of existing social science knowledge?

Research Design: Does the research design have both internal and external validity?

Program Implementation: How does the program actually operate?

Data Collection: Is the necessary data available and reliable?

Measurement: Are the key variables valid and can they be measured reliably?

Analytical Models: Are the data summarized and analyzed by means of appropriate statistical models?

Interpretation of Findings: Are the findings interpreted objectively and do they describe the limitations of the analyses and consider alternative interpretations? Are they placed in the proper policy or programmatic context?

These criteria, of course, are not equally applicable to all evaluations.

Finally, this publication describes a new program to evaluate the evaluations being conducted by the University of Maryland's Welfare Reform Academy.

DOUGLAS J. BESHAROV

Introduction

"The role of social science lies not in the formation of social policy, but in the measurement of its results."
Senator Daniel Patrick Moynihan, 1969[1]

In August 1996, President Clinton signed a welfare reform law, Temporary Assistance for Needy Families (TANF), that could result in fundamental changes in state welfare programs. Building on the extensive array of state waivers his administration and the Bush administration had granted, the new welfare law creates a block grant that caps total federal aid to the states but, in return, allows them much greater flexibility in how they shape their programs.

The new welfare law makes dozens of other changes, including a requirement that a gradually increasing share of state caseloads (including single mothers with young children) must be in work activities; a time limit of five years of benefits (with states free to establish shorter limits, as most have); special residency and education rules for teen mothers; and heightened child support enforcement.[2]

The general public and federal, state, and local officials, along with experts and advocates on the left and right, are eagerly awaiting evidence of the new law's impact (as well as that of the changes resulting from the waivers the federal

government earlier granted to states). Thus, Senator Moynihan's wisdom may soon be demonstrated by a steady stream of research on both the state waiver experiments and the new welfare regime.

Current Evaluations

In the next few years, the major sources of new information will be the continuing evaluations of the 1988 welfare reform law, the Family Support Act (FSA), and of the waiver-based state welfare experiments that followed. Each is described next.

Evaluations of the Family Support Act

The last major attempt to reform welfare was the Family Support Act (FSA) of 1988, which created the Job Opportunities and Basic Skills Training (JOBS) Program within the Aid to Families with Dependent Children (AFDC) Program. The JOBS Program provided AFDC recipients with employment, training, and education-related activities, as well as supportive services.

The FSA also required states to impose work or "participation" mandates on adult recipients. States were to enroll increasing percentages of their "mandatory" AFDC caseload[1] in JOBS activities and to target expenditures on individuals most likely to become long-term welfare recipients. These participation rates were gradually increased from 7 percent in fiscal year 1990 to 20 percent in 1995. For the first time, mothers with children under the age of six were expected to participate and states could extend this require-

ment to those with children as young as one year of age.

The FSA imposed considerably higher participation requirements on two-parent families, who were eligible under the Unemployed Parent (AFDC-UP) Program. Between fiscal years 1994 and 1997, participation was to rise from 40 percent to 75 percent. Moreover, the required activities were generally limited to employment and work activities, with basic education encouraged only for younger recipients who had not finished high school.

The FSA also required states to ensure that teen mothers on welfare finished high school or other education or training programs. (This latter mandate, however, seems initially to have been ignored by most states.)

To ease the transition from welfare to work, the FSA provided up to 12 months of extended child care and Medicaid benefits. The act also required states to adopt even stronger child support enforcement measures, including immediate income withholding, mandatory guidelines for establishing support awards, and periodic review and adjustment of support orders. Finally, it required all states to operate an AFDC-UP program. (At that time, only about half the states provided such aid.)

The major evaluation of the JOBS Program is being conducted by the Manpower Demonstration Research Corporation (MDRC). It includes process, impact, and cost-benefit analyses, as well as special studies related to child well-being and the effectiveness of adult education. A 1995 report describes the program's preliminary, two-year impacts on employment, earnings, and welfare receipt in three sites (Atlanta, Georgia; Grand Rapids, Michigan; and Riverside, California).[2] Its two most notable findings were:

- "Labor force attachment" strategies: Mandatory job search activities followed by work experience or short-term education or training for those who did not find employment reduced welfare receipt 11 percentage points (68 percent for the control group vs. 57 percent for the treatment group).

Monthly welfare payments declined by 22 percent ($276 vs. $216). Employment increased 8 percentage points (34 percent vs. 42 percent) and average monthly earnings increased 26 percent ($226 vs. $285).

- "Human capital development" strategies: Longer-term education and training activities also produced positive results, although not as great as those for labor force attachment approaches. Welfare receipt decreased by 4 percentage points (69 percent vs. 65 percent) and monthly welfare payments declined 14 percent ($285 vs. $247). However, there were no significant impacts on overall employment or earnings.[3]

In addition, the Rockefeller Institute of Government conducted a ten-state, three-year study of the implementation of the JOBS program by state and local governments. The study described how different states used the flexibility under JOBS to implement their programs. The findings covered a wide range of topics, including the design of state JOBS programs, child care and supportive services, case management, and federal participation and targeting requirements. Although it had no impact data, the study concluded that JOBS was a promising approach, but would benefit from additional funding and strong leadership from federal and state lawmakers. In fact, it argued that a major overhaul of the welfare system was not desirable, given the promise of JOBS.[4]

Several demonstration projects are testing JOBS-like services for special populations, such as teen parents, families no longer receiving AFDC, and noncustodial parents of AFDC recipients.

The *Teenage Parent Demonstration* is a randomized experiment being evaluated by Mathematica Policy Research (MPR), Inc. It required teen parents in three cities (Chicago, Illinois; and Newark and Camden, New Jersey) to participate in an education or training activity and then seek employment. Case management and a rich array of services, such

as child care, transportation assistance, and counseling, were also provided. During the two years following intake, participation in school, job training, or employment was 19 percent higher for those in the treatment group than the control group (66 percent vs. 79 percent). However, impacts on employment and earnings were more modest. Employment for the treatment group was 12 percent higher (43 percent vs. 48 percent) and monthly earnings were up 20 percent ($114 per month vs. $137 per month). Monthly AFDC benefits were 7 percent lower ($242 per month vs. $261 per month). However, over 60 percent of the mothers experienced another birth and the birth rate was actually somewhat higher for the treatment group.[5] A study soon to be released will provide the results for a longer follow-up period.

The *Post-Employment Services Demonstration* is another randomized experiment being evaluated by MPR. Although many welfare recipients leave welfare for work, many return. This evaluation tests the impact of providing services, such as assistance with family and social problems, on the likelihood that such families will keep their jobs and stay off welfare.[6]

The *Parents' Fair Share (PFS) Demonstration* is being evaluated by MDRC. It is intended to help reduce the welfare dependency of children by requiring unemployed noncustodial fathers to participate in employment-related activities. The goal of the project is to increase the earnings of these absent parents so that they can provide more financial support for their children.[7]

Evaluations of State Waiver Experiments

In his 1992 State of the Union Address, President Bush encouraged states to seek waivers of federal welfare rules in order to test innovative new programs and policies. President Clinton heightened federal support for the waiver process and also streamlined it.

The actions of both presidents sparked a flurry of state activity. By August 1996, when the new welfare law was

signed, 43 states already had waivers authorizing significant changes in their programs. Waivers authorized states to impose time limits on assistance, strengthen work and training requirements, allow recipients to keep more of their welfare income when they go to work, expand child care and other services for families in work or training, impose requirements and incentives for teen parents to live at home and stay in school, and test many other policies aimed at improving the well-being of needy families with children.[8]

Since the official purpose of these waivers was to allow states to experiment with changes in federal programs, such "experiments" were required to be systematically evaluated. As a result, dozens of large-scale evaluations are in progress, with many initial reports published or soon to be published.

Many of the program changes made through waivers closely resemble the changes that are likely under the new welfare law. Hence, those evaluations that were soundly designed and implemented should contain valuable information to guide state planning and implementation decisions. In fact, for some years to come, the waiver evaluations will likely constitute the best source of information about the probable effects of various programmatic changes.

For example, MDRC recently issued a report on the early experiences in Florida, Vermont, and Wisconsin in implementing time-limited welfare under waivers.[9] Some of its findings were: (1) Implementing time limits without adequate planning poses significant risks to both recipients and the program's credibility. States must carefully assess the need for additional staff and staff training, job training, child care, and management information systems. (2) Communicating the new program rules to recipients is extremely important but can be difficult, especially when several far-reaching changes are implemented simultaneously. It is essential to explain the new policies to recipients clearly and repeatedly. (3) Each state has expanded its JOBS program and is trying to focus on employment, but their approaches vary. For example, some place greater focus on quick job entry and oth-

ers on longer-term approaches. (Impact data from these state demonstrations are only now becoming available.)

In a 1995 study, Pavetti and Duke of the Urban Institute examined waiver programs in five states (Utah, Colorado, Iowa, Michigan, and Vermont).[10] The authors focused on implementation issues raised by attempts to increase JOBS participation substantially and to change the culture of welfare. Their report concluded that states have taken different approaches to reach similar goals.

Some key findings include: (1) Participation rates for work or work-related activities can be substantially raised in a relatively short period of time, but program costs rise. (2) Child care plays an important role in transforming the welfare system into a more work-oriented system. (3) If large numbers of recipients are placed in unsubsidized employment and caseloads decline substantially, those recipients left behind are likely to have multiple barriers to employment. (4) The sanctioning of clients is an important strategy for reforming the welfare system.

In November 1996, Besharov and his colleagues at the American Enterprise Institute collected information from 21 states on the application of health-related rules authorized by the waiver process.[11] Their major findings include: (1) Considerable state interest exists for using financial sanctions or support services to change the behavior of welfare mothers. (2) Most states adopted a program that required recipients to establish their compliance with immunization mandates, sanctioned recipients for noncompliance (either initially or after a warning), and provided "good cause" exemptions. (3) Sanctions for failing to have children immunized ranged from $25 to the entire portion of the mother's grant, usually for as many months as the family was not complying with the requirements. The limited data available suggest that the sanctions were not severely burdensome on the families involved. (4) Neither the monitoring and sanctioning process nor the provision of support services seem to have created an undue or prohibitive administrative burden. (5) Subject

to various methodological and implementation problems, the results of two early evaluations suggest significant increases in immunization rates.

Dozens of additional reports will be issued in the coming months and years. The following is a sampling of what the states are doing.

California's Work Pays Demonstration Project combines benefit reductions with expanded work incentives. An interim impact report suggests that the demonstration has had little effect on employment, earnings, or welfare receipt.[12] However, the implementation analysis concluded that eligibility workers did not effectively communicate to recipients the new rules, especially the financial incentives for working, which may have influenced the absence of effects.

Colorado's Personal Responsibility and Employment Program seeks to reduce welfare dependency by encouraging participation in job training programs and increasing work incentives. A preliminary analysis shows no reduction in welfare receipt and a small increase in employment.[13]

Florida's Family Transition Program combines a two-year time limit with increased incentives for work. An interim implementation study describes some of the early challenges related to implementing a time-limited welfare program. After 15 months, it produced modest increases in employment and earnings, but it had no impact on AFDC payments. In the sixth quarter, the demonstration increased employment 15.4 percent (39.6 percent vs. 45.7 percent) and average earnings by 23.9 percent ($708 per quarter vs. $877). However, there was no statistically significant reduction in AFDC payments. The authors of the Florida report conclude that "FTP's financial work incentives have helped generate an increase in family income without raising welfare spending; however, in part because of the incentives, FTP is not reducing the rate at which people are accumulating months toward the time limit." Because these findings reflect the period before any recipients had exhausted their time-limited benefits, longer-term follow-up is needed to gauge the program's impact.[14]

Georgia's Preschool Immunization Project requires parents to immunize their preschool children. Failure to comply can result in the imposition of financial sanctions. An interim report indicates that the program significantly improved immunization rates.[15] While encouraging, the report's findings should be examined cautiously because only about half of the AFDC families in the treatment and control groups granted permission for evaluators to examine their children's immunization records. Consequently, the data indicating significant increases in all categories of vaccinations may have been distorted by self-selection biases. Families in the treatment group that were in full compliance with the immunization requirements were presumably somewhat more likely to open their children's records to evaluators.

Iowa's Family Investment Plan tests the impact of a social contract that provides time-limited services and tough penalties for those who fail to comply with the agreement. A special report summarizes the findings of a survey of 137 cases whose cash benefits had been terminated for noncompliance: 40 percent experienced an increase in total income, primarily earnings, but nearly half experienced a drop in income, averaging $384.[16] The report is the first in-depth study of what happens to families who are terminated from welfare. Forthcoming reports will assess the demonstration's impact on welfare dependency.

Maryland's Primary Prevention Initiative requires parents to ensure that children meet certain education and preventive health requirements, with sanctions imposed for each child not in compliance with the demonstration's requirements. An interim report found that the demonstration has not had a significant impact on school attendance rates.[17]

Michigan's To Strengthen Michigan Families program was initially designed to test the impact of various work incentives and requirements, but has since been expanded to include other objectives as well, such as increasing immunization rates and requiring minor mothers to live at home. Annual evaluation reports indicate that the state's

welfare reform program has led to modest improvements in employment and earnings, and small reductions in welfare receipt.[18] For families that were receiving assistance when the demonstration started, the program increased employment by 1.2 percentage points and annual earnings by $223, or 7 percent, over the four-year period they were exposed to the new policies.

Minnesota's Family Investment Program combines AFDC and Food Stamps into a single cash grant and expands work incentives. An interim report found that after six months, the program significantly increased welfare receipt and the number of families combining welfare with work, but cautions that it is too early to draw any firm conclusions about the program's impacts.[19]

New York State's Child Assistance Program is a voluntary alternative to AFDC. It provides enhanced work incentives for single-parent families who work and include at least one child covered by a court order for child support from the noncustodial parent. A five-year impact report found a 4 percent reduction in welfare payments and a 20 percent increase in earnings.[20]

Ohio's Learning, Earning, and Parenting demonstration tests the impact of using welfare bonuses and sanctions to enforce school attendance requirements on teen parents. Several reports suggest that the intervention has produced modest impacts on school attendance and completion rates, as well as subsequent employment.[21] At the four-year point, 93 percent of those assigned to the program had qualified for a financial bonus or penalty. Although LEAP increased school enrollment, attendance, and progress through the 11th grade, it did not increase high school graduation for the full sample of teen parents. It also increased employment for this group by nearly 5 percentage points (60 percent vs. 65 percent) in the fourth year of follow-up.

Utah's Single Parent Employment Demonstration is a multifaceted welfare reform demonstration that allows families that appear eligible for AFDC to be diverted from AFDC

through payments that can be up to three times the regular monthly grant. An interim report indicates that about 10 percent of all applicants have been diverted, and only a small number of diverted families have returned to AFDC. Since diversion is but one of many components of the demonstration, it is not possible to know for certain what impact it has had, but it appears to be a promising approach.[22]

Wisconsin's Learnfare program tested the impact of using welfare sanctions to enforce school attendance requirements on all teens receiving welfare, not just teen parents. The evaluation found that the demonstration has had little effect on improving school attendance. It also suggests that, at least in Milwaukee, the program has suffered from substantial implementation failures.[23]

Some evaluations of the waiver experiments will be halted now that a welfare reform law has passed, but many others will continue. The Department of Health and Human Services (HHS) concluded that these experimental programs could provide important information about the implementation and impact of various welfare reform strategies. Hence, in November 1996, it announced support for the continuation of these evaluations (called "State Welfare Reform Evaluation" projects). At least $7.5 million in annual funding will be available. Thirty states, representing 43 demonstration projects, responded by the deadline. As of June 1997, nine states had been funded to continue their evaluations; the other states are eligible to receive planning grants to develop evaluation plans, with $4 million in funds to be awarded through a competitive process.

In addition, HHS is conducting a "Project on State-Level Child Outcomes: Enhancing Measurement of Child Outcomes in State Welfare Evaluations." Twelve states have been selected to participate in a one-year planning project to assess how their existing welfare reform evaluations could be supplemented to provide more in-depth and uniform measures of child outcomes. This will lead to the selection of about five states, with funding of $3.7 million over a three-year

period, to assess the impact of welfare reform on child well-being. The effort is intended to help states expand their data capabilities and to measure and track child outcomes on an ongoing basis. Technical assistance is being provided by Child Trends to assist the states and the federal government in the planning and implementation of this project.

Within the next two or three years, therefore, there will be dozens of reports on a myriad of state experiments, including different approaches to earnings disregards, asset limits, work requirements, sanctions, time limits, transitional benefits, family caps, immunization and school attendance requirements, requirements on teen parents, and child support enforcement.

Related Evaluations

The foregoing discussion has concentrated on research and evaluation studies that focus directly on reforms to the AFDC Program. Many other studies of equal or greater importance are also under way.[24] Some of the most well known are described below.

The New Chance Demonstration provided comprehensive, multi-year education, training, parenting, child care, and other services to young mothers who had children as teenagers and were also high school dropouts. After 42 months, the program raised school attendance and education attainment rates; 52 percent of the treatment group had received a high school diploma or GED compared to just 44 percent of the control group. However, it had no significant impact on many other outcomes, such as employment, earnings, welfare receipt, reading skills, and health status. In addition, the New Chance mothers were more likely to have had another pregnancy than those in the control group, and equally likely to still be on welfare.[25]

Canada's Self-Sufficiency Project offers a temporary earnings supplement to public assistance recipients. Employment increased by 13.1 percentage points (nearly 50 percent) and

average monthly earnings rose $137 (almost 60 percent) in the fifth quarter after enrollment. During this period, the incidence of welfare receipt declined 14 percentage points and average monthly welfare payments fell by $117. However, when the supplemental payment is counted as a government payment, the treatment group is more likely to receive assistance and its payments are higher.[26]

The New Hope Project is a test of a neighborhood-based antipoverty program and welfare alternative operating in Milwaukee.[27] The project is still in its early stages.

Future Evaluations

The major forthcoming evaluations will assess the new welfare regime created by the Temporary Assistance for Needy Families (TANF) Program, which replaced AFDC in 1996. Although states can continue to operate their welfare programs as they have in the past, most observers expect to see fundamental changes. Already, a number of large-scale evaluations have been launched.

National or Multi-State Evaluations

The passage of TANF almost immediately gave rise to research studies proposing to track changes in state welfare systems and estimate their effects on both state welfare agencies and the poor. The studies will employ various sources of data, including existing national surveys, newly established ones, and administrative data sets.

The Census Bureau Survey, mandated by the new federal welfare law, is likely to be the most significant source of national data on welfare reform. The law includes $10 million a year for the Bureau to expand its data collection through the Survey of Income and Program Participation (SIPP).[1] This new "Survey of Program Dynamics" will be built on the data collected in the 1992 and 1993 SIPP panels, thus extending

data collection for these cohorts through 2001. This will provide ten years of longitudinal data on income, patterns of welfare receipt, and the condition of children. Because the panels began three or four years before the enactment of the welfare reform bill, researchers will be able to assess the impact of the bill by comparing this extensive baseline data with data collected after the bill takes effect.

The U.S. General Accounting Office has embarked on a multi-year project to monitor welfare reform, which will include a 50-state overview and an in-depth review of six states. The six-state review will examine how these states structure their new welfare programs, the challenges they encounter, and the outcomes they achieve. The 50-state component will be based on existing data sources and interviews of state officials and others in two counties within each of the six case-study states.

The Urban Institute, in collaboration with Child Trends, Inc., and Westat, Inc., is conducting a multi-faceted study, "Assessing the New Federalism." This $50 million effort will monitor and assess how the devolution[2] of federal responsibility for social welfare programs is being handled by states. It will provide information on the policies, administration, and funding of social programs in all 50 states, with a targeted effort aimed at 13 states. It will include interviews with program managers to determine how they are implementing the new law and surveys of over 50,000 people to collect detailed information about their economic and social circumstances. One of the objectives of the study is to determine the effects of devolution on the well-being of children and families.

The Northwestern University/University of Chicago Joint Center for Poverty Research (also called the Poverty Center) has formed a national advisory panel "to pursue the development of research-ready data from administrative sources to be used for poverty research." It is reviewing administrative data to examine ways of improving its quality so that it can be used for research. In the future, the Poverty Center plans to make grants in support of such research.

The Nelson A. Rockefeller Institute of Government of the State University of New York (Albany) has undertaken "A Study of State Capacity" that will examine the implementation of the new welfare law in order to gauge the capacity of state governments to operate complex social programs. Examining the political, administrative, and programmatic changes in states, it seeks to determine the strengths and weaknesses in their implementation of the law and to identify solutions to the problems encountered. The study will be based on an in-depth review of implementation in seven to ten states, supplemented by a 50-state survey.

Mathematica Policy Research (MPR), Inc., will use existing state administrative data and SIPP data to create a microsimulation model capable of projecting the new welfare law's impact on costs, caseloads, distributional effects, and other outcomes.

U.S. Department of Health and Human Services, through its various organizational units, will fund evaluations on selected subjects, including the Child Care Research Partnership projects, the National Longitudinal Study of Children and Families in the Child Welfare System, the "Welfare Reform Studies and Analyses" project, and several collaborations on topics such as employment stability and immigration.

We also expect researchers to conduct a series of smaller studies based on various large, longitudinal surveys. In addition to the Census Bureau's newly expanded SIPP survey, they will most likely use the Panel Study of Income Dynamics (PSID), begun in 1968, and the National Longitudinal Survey of Youth (NLSY), begun in 1979. Both provide information on annual and monthly income and program participation, and are an important source of data about intergenerational welfare use.

In the past, many important welfare studies have used these surveys. For example, Bane and Ellwood used the PSID to describe the patterns of welfare receipt, including length of time on welfare and the reasons for welfare entry and exit. Their research has enriched our understanding of the het-

erogeneity of the welfare population and aided in the formulation of public policies designed to reduce welfare dependency.[3] Building on their work, Pavetti used the NLSY to analyze time on welfare and the implications for time-limited welfare.[4] The new welfare reforms are sure to increase the use of these databases.

Community- and Neighborhood-Based Evaluations

Several studies are planned to examine the effects of welfare reform at the community or neighborhood level. Unlike broad national or state studies, these studies focus on the law's impact on urban areas, where implementation is likely to pose the greatest challenges and impacts are likely to be the most problematic.

Johns Hopkins University will conduct a "Multi-City Study of the Effects of Welfare Reform on Children," under the leadership of Lindsey Chase-Lansdale, Linda Burton, Andrew Cherlin, Robert Moffitt, and William J. Wilson. It will examine the impact of welfare reform on children in Baltimore, Boston, and Chicago communities. Surveys and administrative data will be used to collect information on families at several points in time, creating a longitudinal database. In addition, children may be tested to provide a fuller assessment of their well-being. These data will be supplemented with ethnographic community studies.

The Manpower Demonstration and Research Corporation (MDRC) will conduct the "Devolution and Urban Change Project" to assess the impact of devolution on families living in economically depressed neighborhoods in four to six large cities. The study will examine changes in the "safety net" in the cities studied and attempt to link agency practices to outcomes for low-income families. The study will use surveys, ethnographic research, administrative records, and other data sources.

Princeton University, through its Office of Population Research, plans to conduct the "Fragile Families and Child

Wellbeing Project." A birth cohort study of unwed parents and their children, the project's principal investigators will be Sara McLanahan, Irwin Garfinkel, and Jeanne Brooks-Gunn. The study will use a longitudinal design to follow, from birth to age four, a new birth cohort of children born to unwed mothers in certain large metropolitan areas. It will provide information on the determinants of child well-being in these families; the factors affecting the involvement of unwed fathers; and the role of extended families, community services, and government policies on these families.

Evaluating the Evaluations

As the foregoing discussion of studies suggests, the next few years will witness a veritable flood of new evaluation reports. The total body of research will be large, complex, and likely to lead to diverse and contradictory findings.

The Need

Many of the evaluations will provide important information about the impact of the new welfare regime on individuals and institutions. They will identify the difficulties and successes that states have had in implementing their reforms, and estimate the impacts of such reforms on the well-being of the poor, especially on their children. These findings, in turn, can help policymakers choose between various program approaches. For example, after MDRC documented the apparent success of "labor force attachment strategies" in reducing welfare caseloads, many states adopted them.

However, many of the evaluations will have such serious flaws that their utility will be sharply limited. For example, because of design and implementation problems, no one may ever know whether New Jersey's "family cap" had any impact on the birth rates of mothers on welfare. (Recently, two outside experts reviewed the evaluation of New Jersey's Family Development Program, which included a family cap

provision. They concluded that there were serious methodological flaws in the evaluation, so an interim report was not released.)

Evaluations can go wrong in many ways. Some have such obvious faults that almost anyone can detect them. Other flaws can be detected only by experts with long experience and high levels of judgment.

The "100-hour rule" demonstrations are an example of the need for the expert review of evaluations. The AFDC-UP Program (abolished by TANF) provided benefits to two-parent families if the principal earner had a significant work history and worked less than 100 hours per month. Because this latter requirement, the so-called "100-hour rule," was thought to create a disincentive for full-time employment, the FSA authorized a set of experiments to alter the rule. Three states (California, Utah, and Wisconsin) initiated demonstrations to evaluate the impact of softening the rule.

Findings from these evaluations suggest that eliminating the rule for current recipients had little impact on welfare receipt, employment, or earnings. But in a recent review, Birnbaum and Wiseman identified many flaws in these studies.[1] First, random assignment procedures were undermined in all three states, so the treatment and control groups were not truly comparable. Second, the states did a poor job of explaining the policy change to the treatment group, limiting its impact on client behavior. Third, some outcomes, such as those related to family structure, were poorly measured.

The proper use of these forthcoming evaluations requires the ability to distinguish relevant and valid findings from those that are not. This does not mean that studies must be perfect in order to be useful. Research projects entirely without flaws do not exist and, arguably, never will.

Almost every evaluation is compromised by programmatic, funding, time, or political constraints. No program has been implemented with absolute fidelity to the original design. No sampling plan has ever been without faults. Some observations and data are missing from every data set. Ana-

lytical procedures are always misspecified to some degree. In other words, evaluation findings are only more credible or less so, and even poorly designed and executed evaluations can contain some information worth noting.

Devolution has further increased the need for careful, outside reviews of research findings. Previously, the federal government required a rigorous evaluation in exchange for granting state waivers, and federal oversight of the evaluations provided some quality control. In keeping with the new welfare law's block-grant approach, the federal government's supervision of the evaluations of state-based welfare initiatives will be curtailed: States are no longer required to evaluate their reforms and, if they do, they can choose any methodology they wish.

Already, there are indications that state discretion under TANF will lead to a proliferation of evaluation designs, some rigorous but many not. As Galster observes, "Many state agencies either lack policy evaluation and research divisions altogether, or use standards for program evaluation that are not comparable to those set by their federal counterparts. The quantity and quality of many state-initiated evaluations of state-sponsored programs may thus prove problematic."[2]

The number of studies purporting to evaluate welfare reform will grow rapidly in the years to come. The challenge facing policymakers and practitioners will be to sort through the many studies and identify those that are credible. It is a task that will be complicated by the volume and complexity of the studies, and the highly charged political atmosphere that surrounds them.

Tension is already building between the conservative lawmakers responsible for crafting the welfare bill and the predominantly liberal scholars involved in monitoring and evaluating it. Many of the researchers now studying the effects of the welfare law were also vocal critics of it. For example, the Urban Institute's $50 million project to assess the "unfolding decentralization of social programs" is being conducted by the

same organization whose researchers, in a highly controversial study, claimed that the new law would push 2.6 million people, including 1.1 million children, into poverty.[3]

This has caused some conservatives in Congress to worry that "pseudo-academic research" will unfairly portray the effects of the welfare overhaul.[4] Undoubtedly, some on the left as well as the right will misuse or oversimplify research findings to their own advantage, but even the perception of bias can limit the policy relevance of research. Good research should be identified, regardless of the ideological stripes of its authors.

Review Criteria

The key issue is the extent to which a discerned fault reduces the credibility of a study. Unfortunately, most policymakers and practitioners are ill-equipped to judge which faults are fatal, especially since they often must act before the traditional scholarly process can filter out invalid results. This is understandable, since assessing evaluation studies often requires both detailed knowledge of the programs involved and a high level of technical expertise.

To help them better assess this research and glean the lessons it offers, this paper also describes and explains the generally accepted criteria for judging evaluations. The criteria, of course, are not equally applicable to all evaluations.

Program "Theory". Underlying every program's design is some theory or model of how the program is conceived to work and how it matches the condition it is intended to ameliorate. An evaluation of the program should describe the underlying social problem it is intended to address and how the causal processes described in the model are expected to achieve program goals. Hence, a critical issue in assessing evaluations is the adequacy of program models.

Special problems are presented by reforms that have several goals. Many of the waiver-based experiments are intended to achieve diverse objectives, such as increasing work

effort and promoting stable families, and, thus, involve multiple interventions. Sometimes the processes can work at cross purposes, placing conflicting incentives on clients. For example, many states have simultaneously expanded earnings disregards and imposed strict time limits. As a result, families that go to work may be able to retain a modest cash grant as a result of the liberalized treatment of earnings, but if they want to conserve their time-limited benefits, they may choose not to take advantage of this incentive. Examination of program theory can reveal such conflicts and identify potential unwanted side effects.

In assessing the adequacy of an evaluation's program theory, questions such as the following should be raised:

- Is there an adequate description of the underlying social problem the intervention is meant to address?
- Does the intervention make sense in light of existing social science theory and previous evaluations of similar interventions?
- Are the hypothesized causal processes by which the reform effort is intended to achieve its goals clearly stated?
- Have potential unwanted side effects been identified?

Research Design. An evaluation's research design is crucial to its ability to answer, in credible ways, substantive questions about program effectiveness. There are two central issues in research design: (1) "*internal validity*," or the ability to rule out alternative interpretations of research findings; and (2) "*external validity*," or the ability to support generalizations from findings to larger populations of interest.

For example, an evaluation that is based solely on measures of client employment levels taken before and after a reform is instituted lacks strong internal validity because any observed changes in employment levels cannot be uniquely attributed to the reform measures. Similarly, an implemen-

tation study of one welfare office in a state system with scores of such offices is of limited external validity because the office studied may not fairly represent all the others.

The effectiveness of a program is measured by comparing what happens when a program is in place to what happens without the program, the "counterfactual." A critical issue is how the evaluation is designed to estimate this difference.

In this respect, randomized experimental designs are considered to be superior to other designs. (Experimental and quasi-experimental designs are discussed in Appendix A.) In a randomized experiment, individuals or families (or other units of analysis) are randomly assigned to either a treatment group to whom the program is given or a control group from whom the program is withheld. If properly conducted, random assignment should result in two groups that, initially, are statistically comparable to one another. Thus, any differences in outcomes between the groups can be attributed to the effects of the intervention with a known degree of statistical precision. Random assignment rules out other possible influences, except for the intervention itself, and therefore has strong internal validity.

Although random assignment is usually the most desirable design, it is not always feasible, especially when a program enrolls all or most of its clientele. Quasi-experimental designs are then employed. They rely on identifying a comparison group with characteristics similar to those of the treatment group, but from another geographic area or time period or otherwise unexposed to the new policy. In some cases, the outcomes of those subject to a new welfare policy may be compared before and after exposure to the new policy.

The major difficulty with quasi-experimental designs is that the members of comparison groups may differ in some unmeasured or undetectable ways from those who have been exposed to the particular program or intervention. Typically, quasi-experimental designs employ statistical analyses to control for such differences, but how well this is done is open to debate. As a result, their internal validity is not as strong

as with randomized experiments. Judging the strength of an evaluation design's internal validity should be an issue at the center of any assessment.

External validity is also crucial for policy purposes. Even an extremely well-designed evaluation with high internal validity is not useful to policymakers if its findings cannot be extrapolated to the program's total clientele.

In large part, an evaluation's external validity depends on how the research population is selected. In many of the waiver-based welfare reform demonstrations, the research sites either volunteered to participate or were selected based on criteria, such as caseload size and administrative capacity, which did not make their caseloads representative of the state's welfare population as a whole. For example, in Florida, sites were encouraged to volunteer in the Family Transition Program. The two sites eventually selected were chosen because they had extensive community involvement and resources that could be committed to the project.[5] In addition, random assignment was phased in so as not to overload the capacity of the new program to provide the promised services. Thus, the findings are unlikely to be representative of what would happen elsewhere in the state (much less the nation), especially if implemented on a large scale.

The evaluations of the new welfare law will employ a variety of research methods, including randomized experiments, quasi-experimental and nonexperimental designs, ethnographic studies, and implementation research. Each has its own strengths and weaknesses. The method used should be linked to the particular questions asked, the shape of the program, and the available resources.

In assessing the adequacy of an evaluation's research design, questions such as the following should be asked:

- Are the impact estimates unbiased (internal validity)? How was bias (or potential bias) monitored and controlled for? Were these techniques appropriate?

- Are the findings generalizable to larger populations (external validity)? If not, how does this limit the usefulness of the findings?

Data Collection. Allen once observed that "adequate data collection can be the Achilles heel of social experimentation."[6] Indeed, many evaluations are launched without ensuring that adequate data collection and processing procedures are in place. According to Fein, "Typical problems include delays in receiving data, receiving data for the wrong sample or in the wrong format, insufficient documentation of data structure and contents, difficulties in identifying demonstration participants, inconsistencies across databases, and problems created when states convert from old to new eligibility systems."[7] Careful data collection is essential for evaluation findings to be credible.

The data used to evaluate the new welfare law will come from administrative records and specially designed sample surveys. In addition, some evaluations may involve the administration of standardized tests, qualitative or ethnographic observations, and other information gathering approaches. Each of these has its own strengths and limitations.

Because administrative data are already collected for program purposes, they are relatively inexpensive to use for research purposes. For some variables, administrative data may be more accurate than survey data, because they are not subject to nonresponse and recall problems, as surveys are.

Some administrative data, however, may be inaccurate, particularly those that are unnecessary for determining program eligibility or benefit amounts. In addition, they may not be available for some outcomes or may cover only part of the population being studied. For example, information related to family structure would only be available for the subset of cases that are actually receiving assistance.

The primary advantage of surveys is that they enable researchers to collect the data that are best suited for the analysis. However, nonresponse and the inability (or unwillingness) of respondents to report some outcomes accurately can

result in missing or inaccurate data. Moreover, surveys can be expensive. Thus, many evaluations use several different data sources.

Unfortunately, evaluation designs are sometimes selected before determining whether the requisite data are available. For example, New Jersey's Realizing Economic Achievement (REACH) program was evaluated by comparing the outcomes of a cohort of similar individuals in an earlier period using state-level data. The evaluator concluded that "shortcomings in the basic evaluation design . . . and severe limitations in the scope and quality of the data available for analysis, make it impossible to draw any policy-relevant conclusions from the results."[8]

Although very few social research efforts have achieved complete coverage of all the subjects from which data are desired, well-conducted research can achieve acceptably high response rates. Several welfare reform demonstrations have been plagued by low response rates, some as low as 30 percent. A high nonresponse rate to a survey or to administrative data collection efforts can limit severely the internal and external validity of the findings. Even when response rates are high, all data collection efforts end up with some missing or erroneous data; adequate data collection minimizes missing observations and missing information on observations made.

The new welfare law significantly complicates data collection and analysis. It will be more difficult to obtain reliable data and data that are consistent across states and over time because states can now change the way they provide assistance. Under past law, both the population and the benefits were defined by federal standards; under the new law, however, the eligible population(s) may vary considerably and the benefits may take many forms (such as cash, noncash assistance, services, and employment subsidies). This will make it more difficult to compare benefit packages, since providing aid in forms other than direct cash assistance raises serious valuation problems.

In addition, states may separate federal and state funds to create new assistance programs. One reason for such a split is that the new law imposes requirements on programs funded with federal dollars, but states have more flexibility with programs financed by state funds. This may have unintended consequences related to data analysis. For example, states may choose to provide assistance with state-funded programs to recipients after they reach the federally mandated time limit. An analysis of welfare spells would identify this as a five-year spell, when in fact welfare receipt would have continued, just under a different program name. Even if states submitted data on their programs, capturing the total period of welfare receipt would require an ability to match data from different programs.

It will be especially difficult to compare events before and after the implementation of the new law, let alone across states and localities. The Census Bureau is already struggling with such issues. For example, until 1996, all states had AFDC programs, but under TANF, they may replace AFDC with one or more state programs, each with its own name. Simply asking survey members about what assistance they receive now requires considerable background work in each state to identify the programs to be included in the survey.

In assessing the adequacy of an evaluation's data collection, questions such as the following should be asked:

- Are the data sources appropriate for the questions being studied?
- Are the data complete? What steps were taken to minimize missing data? For example, for survey-based findings, what procedures were used to obtain high response rates?
- Is the sample size sufficiently large to yield precise impact estimates, both overall and for important subgroups?
- Are the data accurate? How was accuracy verified?

- What statistical or other controls were used to correct for potential bias resulting from missing or erroneous data? Were these techniques appropriate?
- What are the implications of missing or erroneous data for the findings?

Program Implementation. Key to understanding the success or failure of a program is how well it is implemented. Accordingly, a critical issue in evaluating programs is the degree to which they are implemented in accordance with original plans and the nature and extent of any deviations. Descriptive studies of program implementation are necessary for that understanding and for assessing the program's evaluation.

No matter how well-designed and implemented an evaluation may be, if the program was not implemented well, its impact findings may be of little use for policymaking. For example, the impact assessment of Wisconsin's "Learnfare" found that the program had virtually no impact on school attendance, high school graduation, and other related outcomes.[9] The implementation study found that welfare staff experienced difficulties in obtaining the necessary attendance data to ensure school attendance and that penalties for noncompliance were rarely enforced. Thus, the implementation analysis demonstrated that the initiative was never really given a fair test and provided important information to help state decisionmakers fine-tune their program.

In assessing the adequacy of an evaluation of a program's implementation, questions such as the following should be asked:

- Is the program or policy being evaluated fully described?
- Does the evaluation describe how the policy changes were implemented and operated?
- If defective, how did poor implementation affect estimates of effectiveness?

Measurement. Process and outcome variables must have reliable and valid measures. For most evaluations, the principal variables are those measuring program participation, services delivered, and outcomes achieved. An evaluation of a program that attempts to move clients to employment in the private sector clearly needs reliable and valid measures of labor force participation. A program designed to bolster attitudes related to the "work ethic" needs to measure changes in such attitudes as carefully as possible. (Adequate research procedures include advance testing of measurement instruments to determine their statistical properties and validity.)

Especially important are measures of outcomes for which there is no long history of measurement efforts. Because of the half century of concern with measuring labor force participation, such measures have characteristics and statistical properties that are well known. In contrast, social scientists have much less experience measuring such concepts as "work ethic" attitudes, the "well-being" of children, or household and family structures. Many welfare reform efforts now underway are likely to have goals that imply the use of such measures. Whatever measures of such new concepts are used need to be examined carefully in order to understand their properties and validity. (The better evaluations will report in detail about how measures were constructed and tested for reliability and validity.)

In some cases, the intervention itself may affect the measurement of an outcome. For example, Wisconsin's "Learnfare" program requires that AFDC teens meet strict school attendance standards or face a reduction in their benefits. The Learnfare mandate relies on teachers and school systems to submit attendance data. Garfinkel and Manski observe that the program may have changed attendance reporting practices:

> It has been reported that, in some schools, types of absences that previously were recorded as "unexcused" are now being recorded as "excused" or are not being recorded at all. In other schools, reporting

> may have been tightened. The explanation offered is that Learnfare has altered the incentives to record attendance accurately. Some teachers and administrators, believing the program to be unfairly punitive, do what they can to lessen its effects. Others, supporting the program, act to enhance its impact.[10]

In short, program interventions (and sometimes evaluations themselves) can change the measurement of important outcomes.

In assessing the adequacy of an evaluation's process and outcome measures, questions such as the following should be asked:

- Were all appropriate and relevant variables measured?
- Were the measurements affected by response and recall biases? Did subjects misrepresent data for various reasons? Were there Hawthorne effects; that is, did the act of measurement affect the outcome?

Analytical Models. Data collected in evaluations need to be summarized and analyzed by using statistical models that are appropriate to the data and to the substantive issues of the evaluation.

For example, if an important substantive question is whether certain kinds of welfare clients are most likely to obtain long-term employment, the analytical models used must be appropriate to the categorical nature of employment (i.e., a person is either employed or not) and have the ability to take into account the multivariate character of the likely correlates of employment.

Critical characteristics of good analytic models include adequate specification (the variables included are substantively relevant) and proper functional form (the model is appropriate to the statistical properties of the data being analyzed). This is particularly important for quasi-experimental and nonexperimental evaluations.

Developing appropriate analytical models for quasi-experiments has been the subject of much debate. LaLonde[11] and Fraker and Maynard[12] compared the findings from an experimental evaluation of the National Supported Work (NSW) demonstration to those derived using comparison groups drawn from large national surveys that used statistical models purporting to correct for selection biases. The estimated impacts varied widely in the quasi-experimental models and, most importantly, differed from the experimentally derived estimates. LaLonde found that "even when the econometric tests pass conventional specification tests, they still fail to replicate the experimentally determined results."[13]

Not all researchers share these concerns. Heckman and Smith criticize the earlier studies of LaLonde[14] and Fraker and Maynard[15] by arguing that the problem was not with nonexperimental methods per se, but with the use of incorrect models in the analyses.[16] They also claim that the earlier studies did not "utilize a variety of model-selection strategies based on standard specification tests."[17] They add that earlier work by Heckman and Hotz,[18] using the NSW data, "successfully eliminates all but the nonexperimental models that reproduce the inference obtained by experimental methods."[19] Thus, they conclude that specification tests can be a powerful tool in analyzing data from quasi-experimental designs. (The complexity of the statistical issues that arise in some evaluations is clearly beyond the scope of most policymakers.)

In assessing the adequacy of an evaluation's analytical models, questions such as the following should be asked:

- Were appropriate statistical models used?
- Were the models used tested for specification errors?

Interpretation of Findings. No matter how well analyzed numerically, numbers do not speak for themselves nor do they speak directly to policy issues. An adequate evaluation is one in which the findings are interpreted in an even-handed

manner, with justifiable statements about the substantive meaning of the findings. The evaluation report should disclose the limitations of the data analyses and present alternate interpretations.

The data resulting from an evaluation often can be analyzed in several ways, each of which may lead to somewhat different interpretations. An example of how alternative analysis modes can affect interpretations is found in an MDRC report on California's Greater Avenues for Independence (GAIN) program.[20] GAIN is a statewide employment and training program for AFDC recipients, evaluated by MDRC in six counties, ranging from large urban areas, such as Los Angeles and San Diego, to relatively small counties, such as Butte and Tulare. The report presented impact findings for all six counties separately, as well as together, for three years of program operation.

In presenting the aggregate impacts, MDRC gave each county equal weight. As a result, Butte, which represented less than 1 percent of the state's AFDC caseload, had the same weight as Los Angeles, which had almost 34 percent of the state's caseload. Using this approach, MDRC reported that GAIN increased earnings by $1,414 and reduced AFDC payments by $961 over a three-year follow-up period. This gives smaller counties a disproportionate weight in the calculation of aggregate statewide impacts, but was chosen by MDRC because "it is simple and does not emphasize the strong or weak results of any one county."[21] MDRC examined other weighting options. For example, it weighted the impacts according to each county's GAIN caseload. This resulted in an earnings increase of $1,333 and an AFDC payment reduction of $1,087. Although the impact estimates are somewhat similar to those using the first weighting method, the differences are not trivial.

The impacts could also have been weighted based on each county's AFDC caseload, but this option was not discussed. Although Los Angeles county comprised 33.7 percent of the state's AFDC caseload, its share of the GAIN

caseload was just 9.7 percent. In contrast, San Diego county represented just 7.4 percent of the AFDC caseload, but 13.3 percent of the GAIN caseload.[22] As a result, these counties would have very different effects on the aggregate impact estimates, depending on which weighting mechanism is used. Clearly, the interpretation of research findings can be influenced by the ways in which the findings from sites are combined to form overall estimates of effectiveness.

In assessing the adequacy of an evaluation's interpretation of findings, questions such as the following should be asked:

- When alternative analysis strategies are possible, did the evaluation show how sensitive findings are to the use of such alternatives?
- Are alternative interpretations of the data discussed?
- Are important caveats regarding the findings stated?

Conclusion

The coming years will see the publication of many reports evaluating the impact of the new welfare regime. Some of the most significant have been described in this publication. If all goes well, these studies will help policymakers assess the potential consequences of various reform efforts. They also will aid practitioners by identifying implementation challenges and strategies encountered by others implementing similar reform efforts. In fact, because the new welfare reform law gives states unprecedented flexibility in shaping their welfare programs, these various evaluations may constitute, in their totality, the best information on the effects of welfare reform.

These studies will rely on a variety of evaluation designs and they will inevitably vary in their quality and usefulness. There are no perfect evaluations and even poorly executed ones usually contain some findings that are worthwhile. The challenge will be to identify what is useful and apply it to improving programs. These judgments will often require expertise and experience.

To help the public, other scholars, practitioners, and policymakers understand this research and apply its lessons, we have established a blue ribbon committee of experts in evaluation and related social science fields to provide an in-

dependent review of the research on welfare reform; that is, to "evaluate the evaluations."

Each year, the Committee to Review Welfare Reform Research will assess the quality and relevance of the 10 to 25 most significant evaluation studies, identifying those findings that are sufficiently well-grounded to be regarded as credible. It will report its findings in the general media as well as in scholarly and professional journals.

The professional stature of the Review Committee's members is obviously critical to the credibility of its assessments—and to the attention they would receive. Thus, it is composed of experts whose accomplishments in the field of program evaluation and social policy analysis are widely known and respected. At present, the members of the committee are:

Douglas J. Besharov is a resident scholar at the American Enterprise Institute for Public Policy Research and a professor at the University of Maryland School of Public Affairs. He was the first director of the U.S. National Center on Child Abuse and Neglect. He is the author or editor of several books, including *Recognizing Child Abuse: A Guide for the Concerned* (1990), *When Drug Addicts Have Children: Reorienting Child Welfare's Response* (1994), and *Enhancing Early Childhood Programs: Burdens and Opportunities* (1996).

Robert F. Boruch is University Trustee Chair Professor of Education and Statistics at the University of Pennsylvania. A fellow of the American Statistical Association, he has received awards for his work on research methods and policy from the American Educational Research Association, the American Evaluation Association, and the Policy Studies Association. He is the author of nearly 150 scholarly papers and author or editor of a dozen books, including *Randomized Experiments for Planning and Evaluation: A Practical Guide* (1997) and *Evaluation of AIDS Prevention Programs* (1991).

James J. Heckman is Henry Schultz Distinguished Service

Professor of Economics and director of the Center for Social Program Evaluation at the Harris School of Public Policy Studies, University of Chicago. He is co-editor of *Longitudinal Analysis of Labor Market Data* (1985) and numerous scholarly articles on evaluation topics.

Robinson G. Hollister is a professor of economics at Swarthmore College. He has organized and led reviews of the effectiveness of employment and training programs, including *The Minority Female Single Parent Demonstration: New Evidence About Effective Training Strategies* (1990), and was co-editor of *The National Supported Work Demonstration* (1984).

Christopher Jencks is a professor of public policy at the Malcolm Wiener Center for Social Policy, Harvard University. His research areas of interest include social mobility and inequality. He has been a fellow of the American Academy of Arts and Sciences and the National Academy of Social Insurance. His publications include *The Homeless* (1994), *Rethinking Social Policy: Race, Poverty, and the Underclass* (1992), and *Inequality* (1974).

Glenn C. Loury is a professor of economics and director of the Institute on Race and Social Division at Boston University. He has served on several advisory commissions of the National Academy of Sciences and is currently vice president of the American Economic Association. He is author of *One by One, From the Inside Out: Essays and Reviews on Race and Responsibility in America* (1995).

Peter H. Rossi is S.A. Rice Professor Emeritus at the University of Massachusetts (Amherst). He is a past president of the American Sociological Association and has received awards for work in evaluation from the American Evaluation Association, the American Sociological Association, and the Policy Studies Organization. He has authored or coauthored numerous publications, including *Just Punishments: Federal Guidelines and Public Views Compared* (1997), *Feeding the Poor: An Analysis of Five Federal Nutrition Programs* (1997),

Evaluation: A Systematic Approach (1993), and *Down and Out in America: The Origins of Homelessness* (1989).

Isabel V. Sawhill is a senior fellow and holds the Adeline M. and Alfred I. Johnson Chair in Urban and Metropolitan Policy at the Brookings Institution. She served two years as associate director of human resources at the Office of Management and Budget. She is the author or editor of numerous books and articles, including *Welfare Reform: An Analysis of the Issues* (1995) and *Challenge to Leadership: Economic and Social Issues for the Next Decade* (1988).

Thomas C. Schelling is Distinguished Professor at the School of Public Affairs and Department of Economics of the University of Maryland. He is a past president of the American Economic Association. He serves on or chairs committees of the National Academy of Sciences, the Institute of Medicine, and the Social Sciences Research Council. He is the author of eight books and over 120 articles, including *Choice and Consequence* (1984) and *Strategy of Conflict* (1980).

James Q. Wilson is James Collins Professor of Management at the University of California at Los Angeles and a past president of the American Political Science Association. He is the author of numerous books, including *Crime and Human Nature* (1996), *The Moral Sense* (1993), and *Bureaucracy: What Government Agencies Do and Why They Do It* (1991).

Other experts may be added in the future.

For policymakers and program administrators to benefit from the Committee's work, they must be aware of it. Therefore, an extensive dissemination effort will be undertaken through the University of Maryland's Welfare Reform Academy. (See Appendix B.) The Review Committee's assessments will be designed to be understandable by program administrators, policymakers, and the general public, while still being informative to scholars involved in researching welfare reform. They will be published in monographs and used as the basis of articles for scholarly journals, professional

journals, magazines, and newspapers. In addition, we plan to hold quarterly seminars to review the assessments and place them in a policy-relevant context. These sessions will be broadcast on a nationwide TV satellite / downlink network operated by the Welfare Reform Academy.

APPENDIX A

Experimental vs. Quasi-Experimental Designs

Many social welfare programs look successful—to their own staffs as well as to outsiders—because their clients seem to be doing so well. For example, a substantial proportion of trainees may have found jobs after having gone through a particular program. The question is: Did they get their jobs because of the program, or would they have done so anyway? Answering this question is the central task in evaluating the impact of a program or policy. In other words, what would have happened to the clients if they had not been in the program or subject to the policy.

The key task of an impact evaluation is to isolate and measure the program or policy's effects independent of other factors that might be at work, such as local economic conditions, the characteristics of participants, and the quality of the particular project's leadership. To do so, researchers try to establish the "counterfactual"; that is, they try to see what happened to a similar group that was not subject to the program or policy.

Researchers use either experimental or quasi-experimental designs to establish the counterfactual. After describing both approaches, this appendix summarizes their principal strengths and limitations, with illustrations from recent studies.

Experimental Designs

Many social scientists believe that experimental designs are the best way to measure a program or policy's impact. In an experimental design, individuals, families, or other units of analysis are randomly assigned to either a treatment or control group. The treatment group is subjected to the new program or policy, and the control group is not. The experience of the control group, thus, is meant to represent what would have happened but for the intervention.

If properly planned and implemented, an experimental design should result in treatment and control groups that have comparable measurable and unmeasurable aggregate characteristics (within the limits of chance variation). And, from the moment of randomization, they will be exposed to the same outside forces, such as economic conditions, social environments, and other events—allowing any subsequent differences in average outcomes to be attributed to the intervention.

Thus, experimental designs ordinarily do not require complex statistical adjustments to eliminate differences between treatment and control groups. Policymakers can then focus on the implications of findings, rather than "become entangled in a protracted and often inconclusive scientific debate about whether the findings of a particular study are statistically valid."[1] As we will see, the same is not true for quasi-experiments.

In the last 30 years, experimental designs have been used to evaluate a wide range of social interventions, including housing allowances, health insurance reforms, the negative income tax, and employment and training programs.[2] The evaluations of welfare-to-work programs conducted by Manpower Demonstration Research Corporation (MDRC) in the 1980s—which used experimental designs—are widely credited with having shaped the Family Support Act of 1988.[3] Similarly, in the 1990s, Abt Associates evaluated the Job Training Partnership Act (JTPA) program.[4] Its findings, also based

on an experimental design, likewise led to major policy changes.

Experimental designs are not without disadvantages, however. They can raise substantial ethical issues, can be difficult to implement properly, and cannot be used for certain types of interventions.[5]

Ethical issues arise, for example, when the treatment group is subjected to an intervention that may make its members worse off or when the control group is denied services that may be beneficial.[6] In the late 1980s, the state of Texas implemented a random assignment evaluation to test the impact of 12-month transitional child care and Medicaid benefits. When the study began, the treatment group was receiving a benefit (the transitional services) that was otherwise unavailable. Hence, denying the same benefit to the control group did not raise an ethical issue. But a year later, nearly identical transition benefits became mandatory under the Family Support Act. At that point, the control group was being denied what had become part of the national, legally guaranteed benefit package. In the face of complaints, the secretary of Health and Human Services required the control group to receive the benefits, thereby undercutting the experiment.

Sometimes, members of the program staff object to the denial of services built into the experimental design. When they view the experiment as unethical or fear that members of the control group will complain, they sometimes circumvent the procedures of the random assignment process, thus undermining the comparability of the treatment and control groups. This apparently happened, for example, in an evaluation of the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC).[7]

Implementation issues arise in every study.[8] As Rossi and Freeman state, "the integrity of random assignment is easily threatened."[9] For example, "contamination," where control group members are also subjected to all or some of the intervention, was a problem in many of the waiver-based state

welfare reform demonstrations. In many states, the new policies were applicable to all recipients in the state, except for a small control group (usually drawn from a limited number of sites). It was not uncommon for members of these control cases to migrate to other counties and receive the treatment elsewhere. Metcalf and Thornton add:

> Typical forms of distortion or "contamination" include corruption of the random assignment mechanism, provision of the treatment intervention to controls despite proper randomization, other forms of "compensatory" treatment of controls or unintended changes in the control group environment, and distortion of referral flows.[10]

Statistical adjustments cannot always deal successfully with such problems.

In some experiments, members of either the control or treatment groups may not fully understand the rules to which they are subject. For example, in New Jersey's Family Development Program, it appears that many members of the control group believed that the family cap policy applied to them, probably because of the extensive statewide publicity this provision received.[11] Because this may have affected their behavior, it is unlikely that the impact of the demonstration can be determined by comparing the birth rates in the two groups.

In other cases, treatment group members were not made aware of all the changes that affected them. For example, in California, caseworkers did not initially inform recipients of the state's new work incentive provisions. The impact findings suggest that the demonstration had little effect on employment and earnings, but it is unclear whether this is because the policy was ineffective or because the recipients were unaware of its provisions.

Attrition from the research sample and nonresponse are other problems for experiments, although they can also afflict quasi-experiments. First, the research sample may become less representative of the original target population. For example, in the Georgia Preschool Immunization Project,

the evaluator was only able to obtain permission to examine the immunization records from about half of the research sample.[12] If the immunization records for those not included are significantly different from those for whom data are available, nonresponse bias could be a problem. Second, if there is differential attrition between the treatment and control groups, their comparability is undermined and bias may be introduced.[13] This would be especially problematic if the characteristics of those for whom immunization records are available differed systematically in the treatment versus the control group. For example, it may be that those in the treatment group who were in compliance were more likely to open their records to the evaluators—a reasonable assumption, since those who are behind may hesitate for fear of being sanctioned. There was some evidence in the experiment of such differences, based on the characteristics of the clients at the time of random assignment. As a result, the evaluator had to adopt statistical techniques to control for bias.

There is a possibility that the welfare reform intervention, itself, might affect attrition. For example, treatment cases that lose benefits due to a time limit may move to another state to regain assistance, but there would be no corresponding incentive for control cases to do the same.

All of the foregoing implementation problems, of course, apply to quasi-experimental designs as well.

Some interventions are not amenable to randomized experiments. Experimental designs may not be appropriate for interventions that have significant entry effects. For example, a stringent work requirement may deter families from applying for assistance. This effect may not be captured in a random assignment evaluation, because it occurs before the family is randomly assigned.[14] Some reforms may have community-wide effects.[15] For example, they may change the culture of the welfare office, leading caseworkers to treat all clients—treatment and control—differently. Again, this change would not be captured by a simple treatment-control comparison of outcomes.

Furthermore, the random assignment process, itself, can affect the way the program works and the benefits or services available to control group members.[16] For example, in the National JTPA evaluation, extensive outreach was necessary because the assignment of applicants to the control group left unfilled slots in the program. The applicants brought into the program were not the same as those who were effectively "displaced" when assigned to the control group. Thus, the impacts on those who were in the program may not correspond to the impacts on those who would have been in the program in the absence of the demonstration.

Quasi-Experimental Designs

When random assignment is not possible or appropriate, researchers often use quasi-experimental designs. In quasi-experiments, the counterfactual is established by selecting a "comparison" group whose members are not subject to the intervention but are nevertheless thought to be similar to the treatment group.

Participants vs. Nonparticipants. Participants in the program are compared to nonparticipants with similar characteristics on the assumption that both groups are affected by the same economic and social forces. But even though the two groups may appear similar, they may differ in unmeasurable, or difficult to measure, ways. For example, those who voluntarily enroll in a training program may have more desire to find a job than those who do not. Alternatively, those who do not enroll may want to work immediately or in other ways may be in less need of a training program. Both are possibilities that should be considered in interpreting a study's results. Statistical and other methods are sometimes used to control for such "selection effects," but success in doing so has been mixed.

The Urban Institute used this approach to evaluate the Massachusetts Employment and Training Program (ET) choices.[17] It compiled a longitudinal database of information on about 17,000 AFDC recipients, of which half participated

and half did not beyond initial registration and orientation. The nonparticipants served as a comparison group, selected through a statistical procedure that matched the comparison group members to the participant group on several measurable characteristics, including race, sex, age, and family composition. Some characteristics, such as motivation, could not be measured. Although the evaluators attempted to control for "selection bias," the results are still subject to uncertainty.

Comparison Sites. Individuals from other geographic areas are compared to those in the program. This avoids problems of comparing those who volunteer for a program to those who do not (selection effect), but creates other complications. In particular, statistical adjustments are needed for economic and demographic differences between the sites that may influence participant outcomes. This method works best when similar sites are matched, and when the treatment and comparison sites are selected randomly. But if sites can choose whether to receive the treatment or serve as the comparison, selection bias can be a problem. Also, sites that initially appear well matched may become less so, for reasons unrelated to the intervention. (Some events, such as a plant closing, can be especially problematic.)

In one of the rare exceptions to its requirement for an experimental design of waiver-based demonstrations, HHS allowed Wisconsin to evaluate its "Work Not Welfare" demonstration using a comparison site approach. However, Wisconsin selected as treatment sites the two counties that were most interested (and perhaps most likely to succeed) in implementing the demonstration. Besides this important attribute, it turns out that the two counties differed from others in the state on a number of other dimensions (for example, they had lower unemployment rates), thus complicating the analysis. One review of the evaluation plan concludes:

> It is unlikely, however, that matched comparison counties and statistical models will adequately control for the fact that the demonstration counties were preselected. It may not be possible to separate the

> effects of the program from the effects of being in a county where program staff and administrators were highly motivated to put clients to work.[18]

Pre-Post Comparisons. Cohorts of similar individuals from different time periods are compared, one representing the "pre" period and one the "post" period. This also requires statistically controlling for differences between the groups. Using this approach, several studies examined the impact of the AFDC reforms in the 1981 Omnibus Budget Reconciliation Act (OBRA).[19] One problem with pre-post evaluations is that external factors, such as changing economic conditions, may affect the variable of interest, so that the trend established before the new intervention is not as good a predictor of what would have otherwise happened. The evaluation of the 1981 OBRA changes, for instance, had to control for the 1981–1982 recession, which was the worst in 45 years. In fact, there are likely to be many changes and it is difficult to disentangle the impact of a reform initiative from changes that may occur in the economy or in other public policies. For example, studies using this methodology to examine welfare reform in the 1990s would have to control for expansion in the Earned Income Tax Credit (EITC) and the increase in the minimum wage, two important policy changes that could affect labor market outcomes.

Since there may be no more than a few years of data on the "pre" period, the length of follow-up for the "post" period is limited as well. This may be too short a time to test the long-term impact of important policy changes, especially since some changes, such as time limits, may not be fully effective for many years themselves. It also may not be possible to obtain data on some outcomes for the "pre" period, such as measures related to child well-being, particularly if they are not readily available on administrative records. In addition, detailed data on participant characteristics, economic conditions, and other relevant "control" variables are needed. For example, New Jersey's first welfare reform demonstration, the Realizing Economic Achievement (REACH)

program, compared the outcomes of a cohort of recipients subject to REACH to a cohort of similar individuals from an earlier period. Unfortunately, the evaluator concluded that because the limitations with the historical data were so severe, it was not possible to draw any definitive conclusions from the results.[20]

Another way of conducting a pre-post comparison is to examine those participating in the program before and after going through it. The outcomes for the group in the pre-program period serve as the comparison "group" for the same population after the program is implemented. (For example, the employment and earnings of individuals can be compared before and after participation in a training program.)

A major advantage of this design is that it requires data only on program participants. Unfortunately, as Rossi and Freeman note:

> Although few designs have as much intuitive appeal as simple before-and-after studies, they are among the least valid assessment approaches. The essential feature of this approach is a comparison of the same targets at two points in time, separated by a period of participation in a program. The differences between the two measurements are taken as an estimate of the net effects of the intervention. The main deficiency of such designs is that ordinarily they cannot disentangle the effects of extraneous factors from the effects of the intervention. Consequently, estimates of the intervention's net effects are dubious at best.[21]

Comparisons with Secondary Data Sets. Secondary data sets, such as the Census Bureau's Survey of Income and Program Participation (SIPP), or its Current Population Survey (CPS), or other national or state-level data sources, have also been used to develop comparison groups. In such comparisons, a sample of similar persons is identified to represent what would have happened in the absence of the interven-

tion. Many evaluations of the Comprehensive Employment and Training Act (CETA) employed this approach.[22] As with other quasi-experimental methods, selection bias is a problem, because volunteers for the program are compared to nonparticipants. Moreover, complications can arise because the data for comparison group members derived from such secondary sources are generally cruder than for the treatment group.

Time Series/Cross-Sectional Studies. Time series and cross-sectional analyses use aggregate data to compare outcomes either over time or across states (or other political subdivisions), thus attempting to control for variables that can affect the outcome of interest, including a variable that represents the intervention itself. These methods have been commonly used by researchers, but are very sensitive to the specification of the model.[23]

For example, one evaluation of the Massachusetts ET program used time series analysis.[24] A host of explanatory variables were used to reflect the importance of demographic, economic, and policy factors that would be expected to have an impact on the caseload, including a variable to measure the impact of the program being evaluated. The study found that the ET program did not lead to any significant reduction in the welfare rolls in Massachusetts, but the author cautioned:

> Analysis of time series data is often complicated by the fact that many variables tend to change over time in similar ways. For this reason, it may be difficult to separate out accurately the impact of the different factors. Thus, the estimated effects of the explanatory variables may be unstable, changing from one specification of the model to others.[25]

As is evident from the above discussion, a major problem with quasi-experimental designs is selection bias. This arises out of processes that influence whether persons are or are not program participants. Unmeasured differences in personal characteristics, such as the degree of motivation, rather than the program itself, could explain differential outcomes. Sometimes the selection processes are system charac-

teristics, such as differences among welfare offices, which lead some to participate in reform efforts and others not to. Although there are a variety of statistical techniques to correct for selection bias, it is impossible to know with certainty which is most appropriate. And, since these methods result in different estimates, there is always some uncertainty regarding the findings of quasi experiments. Here is how Gary Burtless of the Brookings Institution put it:

> Our uncertainty about the presence, direction, and potential size of selection bias makes it difficult for social scientists to agree on the reliability of estimates drawn from nonexperimental studies. The estimates may be suggestive, and they may even be helpful when estimates from many competing studies all point in the same direction. But if statisticians obtain widely differing estimates or if the available estimates are the subject of strong methodological criticism, policymakers will be left uncertain about the effectiveness of the program.[26]

With experimental designs, such adjustments are unnecessary, since random assignment should equalize the treatment and control groups in terms of both observable and unobservable characteristics.

***

Experimental designs have long been the evaluation method of choice, and should probably be considered first in any evaluation. Many current welfare reform efforts, however, are not amenable to randomized experiments. The new program or policy may cover the entire state, without provision having been made for a control group; the changes made by the state may have affected norms and expectations across the entire community, sample, or agency, so that the control group's behavior was also influenced; and there may be substantial "entry effects," as described above.

Thus, in many circumstances, a quasi-experimental de-

sign will be the preferable approach. Although not nearly as problem-free as experimental designs, they can provide important information about new policies and programs.

The overriding point is that welfare reform efforts should be evaluated as best as possible and the design chosen should be the one most likely to succeed.

Appendix B

The Welfare Reform Academy

In 1997, the School of Public Affairs at the University of Maryland created an academy to help state and local officials, private social service providers, and other interested parties take full advantage of the new welfare reform law. While the law pressures public officials and service providers to make their programs more efficient and better targeted, it also presents an unprecedented opportunity for states to reshape and improve their programs.

The Welfare Reform Academy will provide training in program design, implementation, and evaluation for the Temporary Assistance for Needy Families (TANF), Food Stamp, Medicaid, job training, child care, child welfare, and child support programs. Instruction will cover the following topics:

- understanding the new welfare reform / block grant environment;
- estimating the costs and behavioral consequences of policy decisions;
- implementing programs;
- monitoring programs and evaluating program effects; and
- performance contracting for services.

The academy maintains a small staff of professionals skilled in program management and development. Direct-

ing the academy is Douglas J. Besharov, a member of the faculty who teaches courses on family policy, welfare reform, and the implementation of social policy. Assisting with curriculum development and instruction is Mathematica Policy Research, Inc., one of the most respected public policy research organizations in the nation.

Founded over 25 years ago, Mathematica has expertise on a wide range of social welfare programs. Recent projects include extensive evaluations of the Food Stamp program, analyses of whether health maintenance organizations reduce health care costs, and the development of microsimulation software for modeling the effects of changes in welfare programs.

A West Coast site will be established by the School of Social Welfare at the University of California at Berkeley, whose faculty is distinguished in areas crucial to the academy, including program evaluation, contracting for services, and connecting program support from various funding streams. Members of the Berkeley faculty will conduct some of the training at Maryland, and Maryland faculty will do the same at Berkeley.

Start-up funding for the academy was provided by the Lynde and Harry Bradley and Annie E. Casey Foundations.

The New World of Welfare Reform

States now receive federal welfare funding mainly through an open-ended, but narrowly constrained, categorical program. The new law combines a number of federal income support (TANF), child care, and job training programs into two interrelated block grants. Under the new system, states get more flexibility in return for fixed amounts of federal funding each year.

State and local officials now have much greater freedom to design and implement welfare, job training, child care, and other social welfare programs. For example, the new TANF law seems to allow states to harmonize their welfare and food stamp programs. Such integration could result in the more efficient delivery of services, and might even create more effective services.

A Teaching Academy

Although some states and localities have already begun reshaping their welfare programs, their success will depend on the analytical and decisionmaking skills of agency managers and planners. The Welfare Reform Academy was created to help state and local agencies meet this challenge.

The primary goal of the academy is to create a cadre of managers and planners who can respond fully and creatively to the challenges—and opportunities—presented by the new welfare system of block grants. Through hands-on training in program design, implementation, and evaluation, the academy will equip participants with the skills necessary to reshape social welfare programs according to state and local needs and priorities.

For example, many states and localities may wish to use their welfare block grants to convert traditional welfare programs into workfare or supported work programs. Under workfare, welfare recipients must accept either private or community service jobs in exchange for cash benefits. Under supported work, welfare mothers take private sector jobs and receive benefits in the form of a wage supplement. If designed and implemented properly, these programs might reduce welfare rolls and help recipients become self-sufficient.

Curriculum

Eventually, we expect the academy to train executive and agency officials, legislators, legislative staffers, private social service providers, and other interested parties from across the country.

Training sessions will take place at the University of Maryland School of Public Affairs and the University of California at Berkeley School of Social Welfare. Participants will attend two weeks of intensive training, with a brief break between each one-week session.

Everyone who satisfactorily completes the program will receive a certificate of completion from the University of

Maryland or the University of California. The academy will offer graduate-level education in five areas:

Understanding the New Welfare Reform Environment: What are the specific provisions of the new welfare law, and the choices that states and localities face? This introduction will familiarize participants with the new law and explain critical policy and budget implications. Instructors will explain the policy options that exist under the new system, including ways in which funding streams can be redirected. Participants will explore the possibilities of integrating programs while targeting resources more effectively on specific populations. For example, a state may decide to focus TANF resources on child care services for its low-income population. In addition to covering the range of flexibility states and localities will have, instructors will address the implications of new restrictions included in the legislation.

Estimating Costs and Behavioral Consequences: How to anticipate the likely costs of policy decisions and their impacts on target populations. Some states may be interested, for example, in reducing work disincentives by increasing earnings disregards in income support programs. Or they may seek to create work opportunities for recipients who do not find jobs on their own. The proponents of such policies may be firmly convinced of their merits, but may not fully recognize their cost implications. The academy will teach participants how to estimate the costs of specific proposals, as well as their probable consequences for the populations served, and how to use a cost-benefit approach to program planning. Participants will use micro-simulation software developed by Mathematica to predict how changes in program parameters may affect caseloads. They will also be taught how to develop methods for examining secondary effects, such as how changes in one program can change the cost of others. Thus, the academy will help states minimize the risk of unanticipated costs and other outcomes.

Implementing Programs: How agencies and service providers should implement the new law. The success of a new program depends on how well it is implemented. Instructors will high-

light typical implementation problems and identify useful strategies for overcoming them. Using case examples, the training will focus on effective ways to define program goals, reorganize and motivate staff, redirect resources, delegate responsibility, and assign tasks. Instructors will also discuss how implementation is linked to program monitoring and evaluation.

Monitoring Programs and Evaluating Program Effects: How to monitor programs and assess the actual effects of policy decisions on the well-being and behavior of children and families. Once adopted, new programs must be monitored closely. Instructors will outline the best ways program managers can monitor service delivery, including through specific and quantifiable performance indicators. Participants also will practice using analytical tools for evaluating program effects. New program eligibility rules, administrative arrangements, and program services can change recipient behavior—for better or worse. Determining the impacts of policy changes requires careful evaluation design. The training will cover: (1) how sample design and sampling procedures affect the research questions that can be answered; (2) the types of data that should be collected; (3) how evaluation can be integrated with program and policy implementation; and (4) how resource constraints affect evaluation strategy.

Contracting for Services: How to Make Effective Use of Outside Resources. To plan policy and program changes, transform agency structure and staff practices, or evaluate the effects of reforms, state and local agencies may find it useful to contract with outside vendors for certain services. Contractors may be used to provide basic services, such as job training and child support enforcement; or to supplement internal staff resources; or, in the case of evaluation, to ensure objectivity. Successful contracts require systematic procurement, a clear definition of contractor and agency roles and responsibilities, a sensible degree of contract monitoring, and ongoing communications between agency and contractor. The academy will teach participants how to select

appropriate activities for contracting out and how to evaluate and compare contract proposals. Instructors will also show participants how to attract the kinds of proposals they want, get the most for their money, and avoid common pitfalls of the contracting process. For example, some states and localities may wish to transform traditional services into voucher systems. The training will cover ways to help ensure that such systems work well.

Training will involve assigned reading and homework, class discussions, and individual and team exercises. It will also incorporate various case examples of successful state and local initiatives. During the first stages of the academy's instruction, we will concentrate on the following topics:

1. The History of Welfare and Welfare Reform
2. Current Programs (including AFDC, JOBS, Food Stamp, Medicaid, and Housing)
3. Welfare Caseload Dynamics
4. Eligibility, Income Limits, and Other Requirements
5. Benefit Levels and Interaction Among Programs
6. Earnings Disregards
7. Time Limits
8. Job Training Programs
9. Work Programs
10. Child Care
11. Health Care Coverage
12. Child Support Enforcement
13. Family Caps
14. Learnfare
15. Health-Related Rules
16. Noncitizen Coverage
17. Comprehensive Policy Packages

Notes

Introduction

1. Daniel Patrick Moynihan, *Maximum Feasible Misunderstanding: Community Action in the War on Poverty* (New York, NY: Free Press, 1969), 193.

2. U.S. House of Representatives, Committee on Ways and Means, *1996 Green Book*, 104th Congress, 2nd session, 4 Nov. 1996, 1333-96; Mark Greenberg and Steve Savner, "A Detailed Summary of Key Provisions of the Temporary Assistance for Needy Families Block Grant of H.R. 3734," Center for Law and Social Policy, 13 Aug. 1996.

Current Evaluations

1. The mandatory AFDC caseload excludes those adults who are exempt (because of the age of their children, a disability, or other specified factors) or have good cause for not participating. Nationally, for participation rate purposes under FSA requirements, less than half of the AFDC caseload was considered mandatory in fiscal year 1995.

2. Stephen Freedman and Daniel Friedlander, *The JOBS Evaluation: Early Findings on Program Impacts in Three Sites* (Washington, DC: U.S. Department of Health and Human Services and U.S. Department of Education, Sept. 1995). Associated reports include: (1) Thomas Brock and Kristen Harknett, "Separation versus Integration of Income Maintenance and Employment Services: Which Model Is Best? Findings from a Case Management Experiment," Manpower Demonstration Research Corporation, Jan. 1997. This study examines the impact of integrating income maintenance and employment services in Columbus, Ohio, within the context of a human capital development approach. (2) Stephen Freedman, Daniel Friedlander, Kristen Harknett, and Jean

Knab, "Preliminary Impacts on Employment, Earnings, and AFDC Receipt in Six Sites in the JOBS Evaluation," Manpower Demonstration Research Corporation, Jan. 1997. This study presents two-year findings on the effectiveness of ten programs in six sites. Some emphasized the labor force attachment approach and others the human capital development approach. Six of the programs showed positive impacts on earnings, and nine reduced the average number of months of AFDC receipt. (3) Gayle Hamilton, *The JOBS Evaluation: Monthly Participation Rates in Three Sites and Factors Affecting Participation Levels in Welfare-to-Work Programs* (Washington, DC: U.S. Department of Health and Human Services and U.S. Department of Education, Sept. 1995). This report analyzes the participation patterns of recipients in three sites, including the extent of participation and reasons for nonparticipation. (4) Kristin A. Moore, Martha J. Zaslow, Mary Jo Coiro, Suzanne M. Miller, and Ellen B. Magenheim, *The JOBS Evaluation: How Well Are They Faring? AFDC Families with Preschool-Aged Children in Atlanta at the Outset of the JOBS Evaluation* (Washington, DC: U.S. Department of Health and Human Services and U.S. Department of Education, Sept. 1995). This report provides a description of a range of child outcomes near the beginning of the evaluation in one site, Fulton County, Georgia; it finds that the families in the study are disadvantaged in many ways. Other reports from MDRC describe preliminary impacts in specific sites.

3. Although, after two years, the findings for the labor force attachment approach are more impressive than those of the human capital model, MDRC cautions that education and training programs may initially keep some participants on welfare longer, but are intended to improve the skills necessary to increase self-sufficiency in the long run. Thus, longer follow-up will be necessary to identify the more effective approach. MDRC also cautions that the results should be considered preliminary because the survey data were only available for 39 percent of the full sample (so that sample sizes were small) and follow-up was only for two years, which may not be long enough for the human capital development model to show its full impact.

4. Jan L. Hagen and Irene Lurie, *Implementing JOBS: Progress and Promise* (Albany, NY: The Nelson A. Rockefeller Institute of Government, Aug. 1994).

5. Rebecca Maynard, ed., *Building Self-Sufficiency Among Welfare-Dependent Teenage Parents: Lessons from the Teenage Parent Demonstration* (Princeton, NJ: Mathematica Policy Research, Inc., Jun. 1993).

6. Anu Rangarajan, *Taking the First Steps: Helping Welfare Recipients Who Get Jobs Keep Them* (Princeton, NJ: Mathematica Policy Research, Inc., 1996).

7. Earl Johnson and Fred Doolittle, *Low-Income Parents and the Parents' Fair Share Demonstration: An Early Qualitative Look at Low-*

Income Noncustodial Parents (NCPs) and How One Policy Initiative Has Attempted to Improve Their Ability to Pay Child Support (New York, NY: Manpower Demonstration Research Corporation, 1996).

8. Steve Savner and Mark Greenberg, *The CLASP Guide to Welfare Waivers: 1992–1995* (Washington, DC: Center for Law and Social Policy, 23 May 1995).

9. Dan Bloom and David Butler, *Implementing Time-Limited Welfare* (New York, NY: Manpower Demonstration Research Corporation, 1995).

10. LaDonna Pavetti and Amy-Ellen Duke (with Clemencia Cosentino de Cohen, Pamela Holcomb, Sharon K. Long, and Kimberly Rogers), *Increasing Participation in Work and Work-Related Activities: Lessons from Five State Welfare Reform Projects*, vol. I and II (Washington, DC: The Urban Institute, Sept. 1995).

11. Douglas J. Besharov, Kristina Tanasichuk White, and Mark B. Coggeshall, *Health-Related Welfare Rules* (Washington, DC: American Enterprise Institute for Public Policy Research, Nov. 1996).

12. Rosina M. Becerra, Alisa Lewin, Michael N. Mitchell, and Hiromi Ono, *California Work Pays Demonstration Project: January 1993 through June 1995* (Los Angeles, CA: School of Public Policy and Social Research, UCLA, Dec. 1996).

13. Peggy Cuciti, *Colorado Personal Responsibility and Employment Program: Preliminary Analysis* (Denver, CO: University of Colorado at Denver, Feb. 1997).

14. Dan Bloom, James J. Kemple, and Robin Rogers-Dillon, *The Family Transition Program: Implementation and Early Impacts of Florida's Initial Time-Limited Welfare Program* (New York, NY: Manpower Demonstration Research Corporation, 1997); Dan Bloom, *The Family Transition Program: An Early Implementation Report on Florida's Time-Limited Welfare Initiative* (New York, NY: Manpower Demonstration Research Corporation, Nov. 1995).

15. Larry Kerpelman, David Connell, Michelle Ciurea, Nancy McGarry, and Walter Gunn, *Preschool Immunization Project Evaluation: Interim Analysis Report* (Cambridge, MA: Abt Associates, Inc., 1 May 1996).

16. Thomas Fraker, Lucia Nixon, Jan Losby, Carol Prindle, and John Else, *Iowa's Limited Benefit Plan* (Washington, DC: Mathematica Policy Research, Inc., May 1997).

17. Schaefer Center for Public Policy, *Maryland's Primary Prevention Initiative: An Interim Report* (Baltimore, MD: University of Baltimore, 22 Nov. 1995).

18. Alan Werner and Robert Kornfeld, *The Evaluation of To Strengthen Michigan Families: Final Impact Report* (Cambridge, MA: Abt Associates, Inc., Sept. 1997).

19. Virginia Knox, Amy Brown, and Winston Lin, *MFIP: An Early*

Report on Minnesota's Approach to Welfare Reform (New York, NY: Manpower Demonstration Research Corporation, Nov. 1995).

20. William L. Hamilton, Nancy R. Burstein, August J. Baker, Alison Earle, Stefanie Gluckman, Laura Peck, and Alan White, *The New York State Child Assistance Program: Five Year Impacts, Costs, and Benefits* (Cambridge, MA: Abt Associates Inc., Oct. 1996).

21. Johannes M. Bos and Veronica Fellerath, *Final Report on Ohio's Welfare Initiative to Improve School Attendance: Ohio's Learning, Earning, and Parenting Program* (New York, NY: Manpower Demonstration Research Corporation, Aug. 1997); David Long, Judith M. Gueron, Robert G. Wood, Rebecca Fisher, and Veronica Fellerath, *Three-Year Impacts of Ohio's Welfare Initiative to Improve School Attendance Among Teenage Parents: Ohio's Learning, Earning, and Parenting Program* (New York, NY: Manpower Demonstration Research Corporation, Apr. 1996); Dan Bloom, Hilary Kopp, David Long, and Denise Polit, *Implementing a Welfare Initiative to Improve School Attendance Among Teenage Parents: Ohio's Learning, Earning, and Parenting Program* (New York, NY: Manpower Demonstration Research Corporation, Jul. 1991).

22. Utah Department of Human Services, *Utah Single Parent Employment Demonstration Program: It's About Work*, Preliminary Two Year Report (undated).

23. State of Wisconsin Legislative Audit Bureau, *An Evaluation of Third Semester Effects of the Wisconsin Learnfare Program* (Madison, WI: 1 May 1996).

24. See generally Institute for Research on Poverty, "Monitoring the Effects of the New Federalism: A Conference," *Focus* 18, Special Issue 1996, 12-17.

25. Janet C. Quint, Johannes M. Bos, Denise F. Polit, *New Chance: Final Report on a Comprehensive Program for Disadvantaged Young Mothers and Their Children* (New York, NY: Manpower Demonstration Research Corporation, Jul. 1997); Janet Quint, Denise Polit, Hans Bos, and George Cave, *New Chance: Interim Findings on a Comprehensive Program for Disadvantaged Young Mothers and Their Children* (New York, NY: Manpower Demonstration Research Corporation, Sept. 1994).

26. David Card and Philip Robins, *Do Financial Incentives Encourage Welfare Recipients to Work? Initial 18-Month Findings from the Self-Sufficiency Project* (Ottawa, Ontario: Social Research and Demonstration Corporation, Feb. 1996).

27. Dudley Benoit, *The New Hope Offer: Participants in the New Hope Demonstration Discuss Work, Family, and Self-Sufficiency* (New York, NY: Manpower Demonstration Research Corporation, 1996).

Future Evaluations

1. The Survey of Income and Program Participation (SIPP), pro-

duced by the Census Bureau, collects monthly information on about 20,000 households for a period of two-and-a-half years. It collects detailed information regarding employment, income, and participation in social programs. Because it is longitudinal, it is particularly useful for analyzing changes in income and program participation over time. The Census Bureau's Current Population Survey (CPS), the primary source of information on income and poverty in the United States, also may be used by researchers to analyze the impact of the new welfare law. Based on a sample of 60,000 households surveyed each March, it collects data on the demographic and economic characteristics of the sample individuals and households in the preceding year.

2. Devolution is the assignment of planning and decisionmaking responsibilities to lower levels of government or even to communities.

3. Mary Jo Bane and David Ellwood, *Welfare Realities: From Rhetoric to Reform* (Cambridge, MA: Harvard University Press, 1994).

4. LaDonna Pavetti, "Who Is Affected by Time Limits?" in *Welfare Reform: An Analysis of the Issues*, ed. Isabel V. Sawhill (Washington, DC: The Urban Institute, 1995).

Evaluating the Evaluations

1. See Matthew Birnbaum and Michael Wiseman, "Extending Assistance to Intact Families: State Experiments with the 100-Hour Rule," *Focus* 18, Special Issue 1996, 38–41.

2. George Galster, "The Challenges for Policy Research in a Changing Environment," in *The Future of the Public Sector* (Washington, DC: The Urban Institute, Nov. 1996).

3. Sheila Zedlewski, Sandra Clark, Eric Meier, and Keith Watson, *Potential Effects of Congressional Welfare Reform Legislation on Family Incomes* (Washington, DC: The Urban Institute, 26 Jul. 1996).

4. John Harwood, "Think Tanks Battle To Judge the Impact of Welfare Overhaul," *The Wall Street Journal*, 30 Jan. 1997, A1.

5. Dan Bloom, *The Family Transition Program: An Early Implementation Report on Florida's Time-Limited Welfare Initiative* (New York, NY: Manpower Demonstration Research Corporation, Nov. 1995).

6. Jodie Allen, "An Introduction to the Seattle/Denver Income Maintenance Experiment: Origins, Limitations, and Policy Relevance," *Proceedings of the 1978 Conference on the Seattle and Denver Income Maintenance Experiments* (Olympia, WA: Department of Social and Health Services, 1979), 18.

7. David J. Fein, "Waiver Evaluations: The Pitfalls—and the Opportunities," *Public Welfare*, Fall 1994, 27.

8. Paul Decker, *REACH Welfare Initiative Program Evaluation: Estimating the Effects of the REACH Program on AFDC Receipt* (Princeton, NJ: Mathematica Policy Research, Inc., Aug. 1991), 1.

9. State of Wisconsin Legislative Audit Bureau, *An Evaluation of Third Semester Effects of the Wisconsin Learnfare Program* (Madison, WI: 1 May 1996).

10. Charles Manski and Irwin Garfinkel, "Introduction" in *Evaluating Welfare and Training Programs*, ed. Charles Manski and Irwin Garfinkel (Cambridge, MA: Harvard University Press, 1992), 8.

11. Robert LaLonde, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76, Sept. 1986, 604–20.

12. Thomas Fraker and Rebecca Maynard, "Evaluating Comparison Group Designs with Employment-Related Programs," *Journal of Human Resources* 22, Spring 1987, 194–227.

13. Robert LaLonde , "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76, Sept. 1986, 617.

14. *Ibid*.

15. Thomas Fraker and Rebecca Maynard, "Evaluating Comparison Group Designs with Employment-Related Programs," *Journal of Human Resources* 22, Spring 1987.

16. James J. Heckman and Jeffrey A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9, Spring 1995, 85–110.

17. *Ibid*, 91.

18. See James J. Heckman and Joseph V. 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, Dec. 1989, 862–74.

19. James J. Heckman and Jeffrey A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9, Spring 1995, 91.

20. James Riccio, Daniel Friedlander, and Stephen Freedman, *GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program* (New York, NY: Manpower Demonstration Research Corporation, Sept. 1994).

21. *Ibid*, 125.

22. *Ibid*, 10.

Appendix A

1. Gary Burtless, "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives* 9, Spring 1995, 69.

2. David Greenberg and Mark Shroder, *Digest of Social Experiments* (Madison, WI: Institute for Research on Poverty, University of Wisconsin, 1991).

3. Erica Baum, "When the Witch Doctors Agree: The Family Sup-

port Act and Social Science Research," *Journal of Policy Analysis and Management* 10, Fall 1991, 603–15.

4. Larry L. Orr, Howard S. Bloom, Stephen H. Bell, Winston Lin, George Cave, Fred Doolittle, *The National JTPA Study: Impacts, Benefits, and Costs of Title II-A* (Bethesda, MD: Abt Associates Inc., Mar. 1994).

5. See generally Peter H. Rossi and Howard E. Freeman, *Evaluation: A Systematic Approach* 5, 5th ed. (Newbury Park, CA: SAGE Publications, Inc., 1993).

6. Of course, whether a particular intervention makes someone better off or worse off cannot be determined *a priori*.

7. Michael J. Puma, Janet DiPietro, Jeanne Rosenthal, David Connell, David Judkins, and Mary Kay Fox, *Study of the Impact of WIC on the Growth and Development of Children. Field Test: Feasibility Assessment*. Final Report: Volume I (Cambridge, MA: Abt Associates Inc., 1991).

8. Anne Gordon, Jonathan Jacobson, and Thomas Fraker, *Approaches to Evaluating Welfare Reform: Lessons from Five State Demonstrations* (Princeton, NJ: Mathematica Policy Research, Inc., Oct. 1996.)

9. Peter H. Rossi and Howard E. Freeman, *Evaluation: A Systematic Approach* 5, 5th ed. (Newbury, CA: SAGE Publications, Inc., 1993).

10. Charles E. Metcalf and Craig Thornton, "Random Assignment," *Children and Youth Services Review* 14, 1992, 152.

11. Peter Rossi, "What the New Jersey Experiment Results Mean and Do Not Mean," in *Addressing Illegitimacy: Welfare Reform Options for Congress* (Washington, DC: American Enterprise Institute for Public Policy Research, 11 Sept. 1995).

12. Larry C. Kerpelman, David B. Connell, Michelle Ciurea, Nancy McGarry, and Walter Gunn, *Preschool Immunization Project Evaluation: Interim Analysis Report* (Cambridge, MA: Abt Associates Inc., 1 May 1996).

13. Rossi and Freeman explain: "Although randomly-formed experimental and control groups are `statistically equivalent' at the start of an evaluation, non-random processes may threaten their equivalence as the experiment progresses. Differential attrition may introduce differences between experimentals and controls. In the income maintenance experiments, for example, families in the experimental groups who received the less generous payment plans and families in the control groups were more likely to stop cooperating as subjects." Peter H. Rossi and Howard E. Freeman, *Evaluation: A Systematic Approach* 5, 5th ed. (Newbury, CA: SAGE Publications, Inc., 1993).

14. Robert Moffitt, "Evaluation Methods for Program Entry Effects," in *Evaluating Welfare and Training Programs*, ed. Charles Manski and Irwin Garfinkel (Cambridge, MA: Harvard University Press, 1992), 231–52.

15. Irwin Garfinkel, Charles F. Manski, and Charles Michalopoulos, "Micro Experiments and Macro Effects," in *Evaluating Welfare*

and Training Programs, ed. Charles Manski and Irwin Garfinkel (Cambridge, MA: Harvard University Press, 1992), 253–76.

16. James J. Heckman and Jeffrey A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9, Spring 1995, 85–110.

17. Demetra S. Nightingale, Lynn C. Burbridge, Douglas Wissoker, Lee Bawden, Freya L. Sonenstein, and Neal Jeffries, *Experiences of Massachusetts ET Job Finders: Preliminary Findings* (Washington, DC: The Urban Institute, 1989).

18. Anne Gordon, Jonathan Jacobson, and Thomas Fraker, *Approaches to Evaluating Welfare Reform: Lessons from Five State Demonstrations* (Princeton, NJ: Mathematica Policy Research, Inc., Oct. 1996), 23.

19. Research Triangle Institute, *Final Report: Evaluation of the 1981 AFDC Amendments* (Research Triangle Park, NC: Research Triangle Institute, 15 Apr. 1983); Ira Muscovice and William J. Craig, "The Omnibus Budget Reconciliation Act and the Working Poor," *Social Service Review* 58, Mar. 1984, 49–62; and U.S. General Accounting Office, *An Evaluation of the 1981 AFDC Changes: Initial Analyses* (Washington, DC: U.S. Government Printing Office, 1984).

20. Paul Decker, *REACH Welfare Initiative Program Evaluation: Estimating the Effects of the REACH Program on AFDC Receipt* (Princeton, NJ: Mathematica Policy Research, Inc., Aug. 1991).

21. Peter H. Rossi and Howard E. Freeman, *Evaluation: A Systematic Approach* 5, 5th ed. (Newbury, CA: SAGE Publications, Inc., 1993), 250. "In general, then, simple before/after reflexive designs provide findings that have a low degree of credibility. This is particularly the case when the time elapsed between the two measurements is appreciable—say, a year or more—because over time it becomes more and more likely that some process occurring during the time period may obscure the effects of the program, whether by enhancing them or by diminishing them." (p. 343).

22. Burt Barnow, "The Impact of CETA Programs on Earnings: A Review of the Literature," *Journal of Human Resources* 22, Spring 1987, 157–93.

23. Peter H. Rossi and Howard E. Freeman, *Evaluation: A Systematic Approach* 5, 5th ed. (Newbury Park, CA: SAGE Publications, Inc., 1993).

24. June O'Neill, *Work and Welfare in Massachusetts: An Evaluation of the ET Program* (Boston, MA: Pioneer Institute for Public Policy Research, 1990).

25. *Ibid*.

26. Gary Burtless, "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, Spring 1995, 72.