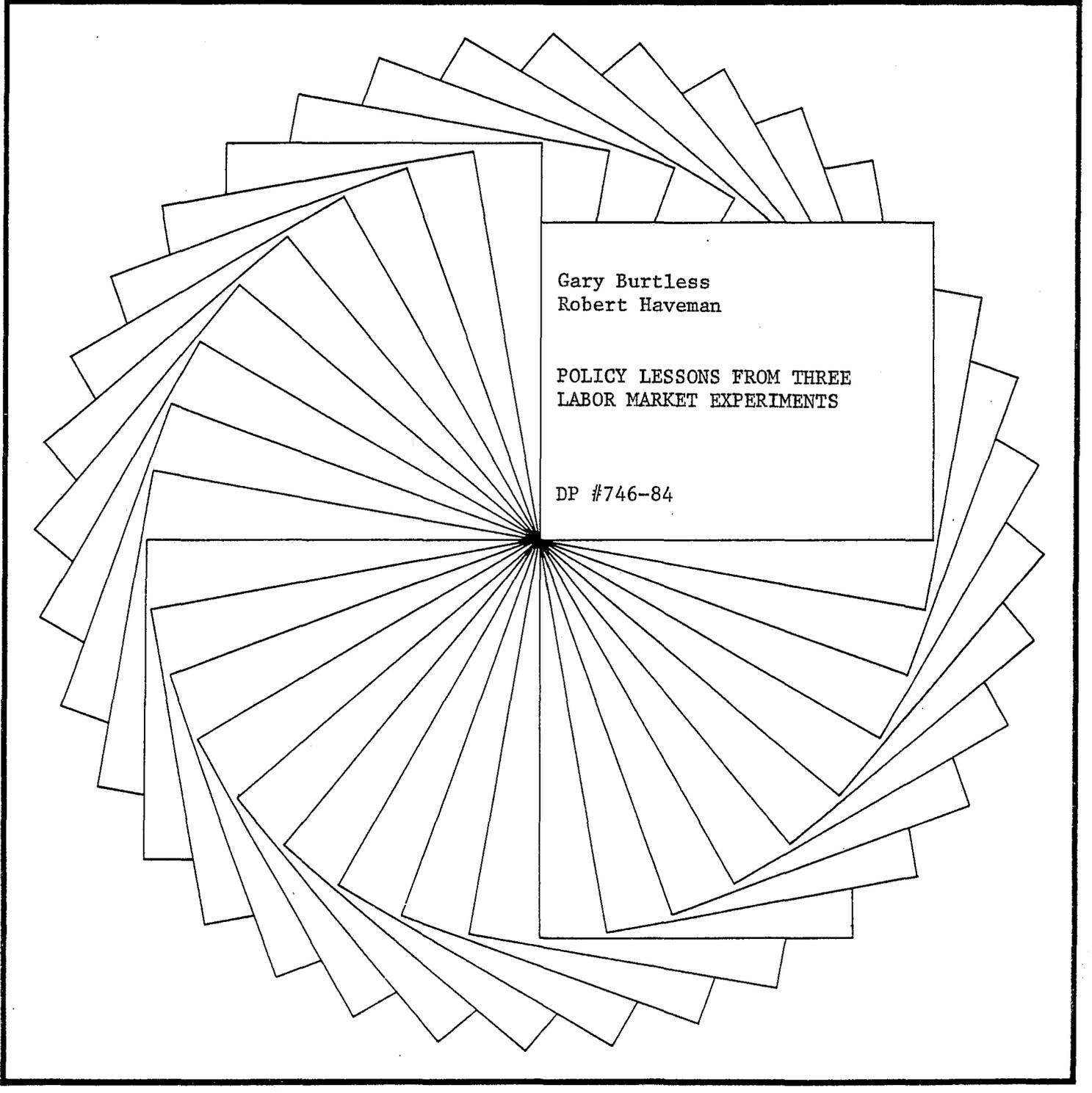

IRP Discussion Papers

A graphic consisting of a fan of approximately 20 rectangular papers fanned out from a central point. The papers are arranged in a semi-circle, with the top edge of the fan curving upwards. A white rectangular box is superimposed on the right side of the fan, containing text. The entire graphic is enclosed in a black rectangular border.

Gary Burtless
Robert Haveman

POLICY LESSONS FROM THREE
LABOR MARKET EXPERIMENTS

DP #746-84

Policy Lessons from Three Labor Market Experiments

Gary Burtless
The Brookings Institution

Robert H. Haveman
Institute for Research on Poverty and
Department of Economics
University of Wisconsin-Madison

March 1984

This paper was commissioned by the National Council on Employment Policy, Washington, D.C. We are grateful to Elise Bruml, David Greenberg, Richard Nathan, Larry Orr, Philip Robins, and Felicity Skidmore for helpful comments. The opinions expressed are solely our own and do not necessarily represent those of the National Council, the Brookings Institution, or the Institute for Research on Poverty.

ABSTRACT

This paper discusses the design, administration, and findings of three social experiments or demonstrations conducted in the past decade: the Seattle-Denver Income Maintenance Experiment (SIME-DIME), the National Supported Work Demonstration (NSWD), and the Employment Opportunity Pilot Project (EOPP). All three projects tested forms of public intervention designed to affect the work effort and incomes of low-income families. In all cases, the effect of the intervention on the earnings of disadvantaged groups was a central issue. The success of the projects was mixed, with respect both to adequacy of design and to the importance and reliability of the findings obtained. For some interventions and target groups, statistically significant earnings effects were found; for other interventions and groups, no impact could be reliably detected. Both SIME-DIME and NSWD contributed substantially to the appraisal of the particular intervention on which they focused. EOPP was less carefully designed and, in part for political reasons, was poorly executed. Its findings are less reliable, and consequently less important, than those from the other two projects.

The concluding section of the paper attempts to describe some fundamental lessons that can be learned from the projects. That section considers the larger political and economic implications of the research. In addition, it suggests several criteria for evaluating proposed demonstrations or experiments before they are initiated.

Policy Lessons from Three Labor Market Experiments

Social experimentation began in earnest when the New Jersey Negative Income Tax Experiment was launched in 1967. For the next fourteen years, government agencies and philanthropic organizations sponsored a wide variety of experiments and demonstrations involving innovations in social policy; none were more important than those concerning the controversial issues of income support and work. In this paper we consider three of the most important social policy experiments: the Seattle-Denver Income Maintenance Experiment, the National Supported Work Demonstration, and the Employment Opportunity Pilot Project. These projects have yielded findings of broad significance to social policy, though the significance of their findings is only dimly perceived by policy makers and interested scholars. Our purpose in this review is briefly to describe the experiments and state the main policy conclusions that can be drawn from them. In the final section we discuss some conclusions about the effects and value of social experiments in general.

THE SEATTLE-DENVER EXPERIMENT

The Seattle-Denver experiment was the largest and most comprehensive of the NIT experiments. It was begun in Seattle in 1970 and in Denver in 1971 under contracts between the states of Washington and Colorado and the U.S. Department of Health, Education, and Welfare. The experiment was administered by Mathematica Policy Research, an organization that had already gained valuable administrative experience running the New Jersey experiment. The Stanford Research Institute designed the experiment and

was given major responsibility for evaluating it. There is no doubt that the Seattle-Denver project was the best run of the NIT experiments, and it was the most thoroughly studied.

Approximately 4800 families were enrolled in the experiment, and families assigned to experimental NIT plans were potentially eligible for payments for a period of either three or five years.¹ To be eligible for enrollment, families had to contain at least one able-bodied, nonaged adult. If only a single adult was present, the family was also required to have one or more dependent children. The sample enrolled in the experiment consisted of lower- and middle-income black, white, and Hispanic families with either one or two parents present. While participation was restricted to residents of Seattle and Denver, families could continue to participate if they moved out of those cities.

The experiment had two main goals, both of which were reflected in its rather elaborate design. The first was to determine the effect of alternative NIT plans on the work behavior of the poor. The second was to test the feasibility and effectiveness of educational vouchers aimed at low-income workers.

The NIT Plans

The idea behind a negative income tax (or NIT) is fairly well known and will not be discussed in detail here. In its simplest form an NIT offers a guaranteed monthly or annual income to a family that has no other income of its own. This amount varies depending on the number of persons in the family and was systematically varied in the experiment to measure the impact of higher or lower income support levels. If a family

received income from nonexperimental sources, such as wage earnings, interest, or public transfers, the monthly NIT payment was reduced in proportion to the amount of other income received. As income from other sources rose, the NIT payment was reduced by an amount determined by the program's tax (or benefit reduction) rate. The tax rate was also systematically varied in the experiment. When income from other sources was sufficiently high that the benefit reduction exactly offset the income guarantee--at a point known as the break-even--payments under the NIT ceased. An NIT's break-even level is algebraically determined by its guarantee and tax rate. As the guarantee level rises, the break-even also rises; as the tax rate rises, the break-even level declines.

Both theory and common sense suggest that the transfer scheme just described will affect work effort. Those who receive payments will have more income, so the necessity for earned income falls. Because payments are reduced as earned income rises, the reward for work is also affected. Under a benefit reduction rate of 70 percent, for example, a recipient who earns an additional dollar loses \$0.70 in NIT benefits, and the net increase in income is only \$0.30. The Seattle-Denver experiment tested eleven NIT plans with income guarantees ranging from slightly below to about 40 percent above the poverty threshold and tax rates ranging from about 50 to 80 percent. With this range of tested guarantees and tax rates, the designers hoped to detect the impact of a meaningful array of plans. In retrospect, we can criticize the designers for their conservative assessment of what constituted a "meaningful" range of tax rates. The policy debate since 1977, and especially since 1981, has shown that tax rates in excess of 90 percent or even at 100 percent are well inside

the policy-relevant range, although this certainly was not foreseen when the experiment was designed.

The random assignment of families or individuals to alternative treatments--or no treatment at all--is what gives social experimentation its unique advantage as a tool for policy analysis. With only a few modest and believable statistical assumptions, it is possible for the analyst of experimental data to establish a definite cause-and-effect relationship between treatment variations and observed outcomes. The direction and precise magnitude of the relationship can be established with known levels of statistical confidence. In the case of the Seattle-Denver experiment, families were randomly assigned to one out of the eleven NIT plans or to control status. A family enrolled in one of the NIT plans was eligible to receive NIT grants if its income was below the plan's break-even. A family in the control group was not eligible to receive these experimental transfers but could continue to receive any nonexperimental transfers for which it remained eligible. The effect of the NIT plans on work behavior can be reliably determined simply by statistically comparing the work effort of individuals enrolled in the various plans and in the control group.

The work-effort findings from the Seattle-Denver experiment have been summarized in a final report issued in 1983 by the Department of Health and Human Services. Briefly, the report shows that the NIT plans caused substantial reductions in labor market activity, particularly for persons enrolled in longer duration (five-year) plans and for women. By "substantial" we mean that prime-aged men reduced their annual hours of work by 9 or 10 percent; that their spouses reduced annual hours by 17 to

20 percent; and that single women heading families reduced annual hours by more than 20 percent--perhaps by as much as 28 to 32 percent.² These reported work reductions are large enough to cause alarm among conservatives already opposed to an NIT and even among centrists with no strong opinions about the desirability of an NIT.

Taken by themselves, however, the work reductions just reported have almost nothing to tell us about the desirability or feasibility of enacting an NIT. The work reductions appear to be fairly substantial, but the work disincentive provided by the tested plans was also quite substantial, larger in fact than that which would be provided under most proposed NIT plans. The Seattle-Denver plans tested an average income guarantee of 115 percent of the poverty threshold and a marginal tax averaging about 50 percent. In addition, the experiment provided rebates for state, federal, and payroll taxes on earned income. About 80 percent of enrolled families faced a break-even level that was more than one-and-a-half times the poverty threshold, and 50 percent faced a break-even more than twice the poverty level (that is, above \$19,600 for a family of four in 1982 dollars). By contrast, the combined income guarantee provided by AFDC and food stamps is now below the poverty level in most states, and the break-even level for AFDC is below the poverty level in all but 15 states.³

Even so, the labor supply findings from Seattle-Denver were considered sufficiently important to affect the welfare reform proposals submitted by the Carter administration.⁴ The reason was quite simple. The results showed quite convincingly that the work incentive provided by

an NIT's low marginal tax rate was more than offset by the work disincentive effects caused by higher overall transfers. For example, simulations based upon the Seattle-Denver results demonstrated that replacement of the current welfare and food stamp programs with a national NIT that has a guarantee equal to three-quarters of the poverty line and a marginal tax rate of 50 percent would reduce aggregate labor supply in two-parent families by about 1 percent. Labor supply in two-parent families with annual incomes below \$5,000 would be reduced by more than 8 percent.⁵ Although we do not find these estimates discouraging by themselves, they contain an implication that is dispiriting to policy makers who wish simultaneously to support incomes and increase the self-reliance of needy families. According to the Seattle-Denver estimates, under the NIT plan just described it would cost the government \$1.79 in transfer outlays to raise the net income of poor two-parent families by \$1.00. In other words, 44 percent of the net program costs of the NIT would be "consumed" by breadwinners in the form of leisure. (The net program cost of the NIT is the amount by which NIT transfers exceed those now paid under the welfare and food stamp programs.)

Another important--though at first glance perverse--result from the experiment was that lowering work incentives in transfer programs by raising their marginal tax rates (holding the guarantee constant) served to increase aggregate work effort. For example, if the tax rate in the NIT just described were raised from 50 to 70 percent, the Seattle-Denver results indicated that aggregate work effort would rise by 1 percent.⁶ The result is attributable to the fact that while increases in marginal

tax rates may indeed reduce the aggregate work effort of those who continue to receive transfers, that effect is more than outweighed by the increases in work effort that occur among those who lose benefits altogether. (Recall that a rise in the marginal tax rate with a constant guarantee causes a fall in the break-even and hence a reduction in the number of transfer recipients.)

If one's sole objective is to increase work effort, the recent increases in AFDC tax rates might conceivably be justified by findings of the Seattle-Denver experiment.⁷ This conclusion, however, rests on the premise that the main objective of transfer policy is to encourage work effort. In fact, the primary objective of an NIT is to protect the living standards of people who would otherwise be destitute, and to do so in an equitable and efficient way. The contribution of the NIT program to this objective, it should be noted, has received only slight attention in the hundreds of research reports filed on the NIT experiments--this in spite of the fact that the tested NIT plans were potentially quite effective in attaining that goal. Nevertheless, the Seattle-Denver experiment has played the useful role of overturning the notion, especially popular among economists and idealistic reformers, that lower marginal tax rates are automatically associated with a greater stimulus to work.

Education and Job Training

The second objective of the experiment was to test the effectiveness of issuing education and training vouchers to low-income breadwinners. Families in the experiment were randomly assigned to one of three employment-training programs or to control status.⁸ All three of the

labor market programs provided a structured course of manpower counseling to help participants decide on an appropriate strategy of employment, education, and training. This course was voluntary, informational in content, and nondirective (that is, participants were not encouraged to pursue any particular course of action). One of the tested programs offered no service beyond this counseling. The other two offered subsidies to pay for some or all of the direct costs of schooling or training.⁹ Two levels of voucher subsidy were tested. In the more generous plan, 100 percent of direct schooling and training costs were reimbursed by the experiment. In the other plan, only 50 percent of costs were reimbursed. Participants could use their vouchers to pay for any form of education or training they chose, so long as it was at least tangentially related to improving their future job prospects.

The purpose of the vouchers was to encourage eligible breadwinners to invest in worthwhile training and education, which according to human capital theory should have improved participants' employability and future earnings. Participation in the program was reasonably high. About one-fifth of family heads in two-parent families used the 50 percent vouchers, and over one-third used the 100 percent vouchers. About one-third of single mothers eligible for the 50 percent vouchers used them, as did nearly one-half of those eligible for the 100 percent vouchers. Not surprisingly, much of the subsidy went to pay for schooling that would have been obtained in the absence of the program. Most of the subsidies paid for attendance in formal academic programs, such as those run by community colleges, rather than for technical training. The more generous subsidy program succeeded in encouraging

extra investment in formal schooling, with the average increase in schooling equaling about one-half an academic quarter among men eligible for the subsidies (relative to those not eligible) and about one to one-and-one-half extra quarters among eligible women.¹⁰

The interesting finding from this experiment is the complete lack of evidence that the increased investment in schooling by participants led to any payoff in the job market. On the contrary, persons eligible for vouchers--in comparison to control-group members--suffered short-term reductions in wage rates, earnings, and employment during the initial phase of their eligibility. This could, of course, be attributed to temporary investment in schooling to achieve higher earnings later. Yet those eligible for vouchers never showed consistent earnings gains over the entire six-year span for which information is available, a period which includes a fairly lengthy spell after participants had completed their schooling.¹¹

An overall explanation for this result is that the vouchers induced significant short-term reductions in work effort and work intensity by subsidizing an alternative use of time--enrollment in formal schooling. After the training was completed, participants' earnings failed to rise above the level observed in the control group because of the amount and character of extra schooling obtained. The amount of extra schooling was on average very small, and it was apparently not particularly relevant to the participants' labor market situation. A second explanation concerns the effect of a rather poor and generally deteriorating labor market on the earnings potential of those who reduced (or ceased) their work in order to obtain additional schooling. In a tight labor market, the

returns to work experience and job-keeping may exceed those to increased schooling. It is difficult to make training pay off if there are few jobs available.

Employment and training programs for the poor are sometimes criticized for being too rigid, too bureaucratic, too paternalistic, and too insensitive to the special needs of different clients. The experimental test of manpower vouchers in Seattle and Denver shows that completely decentralized decision-making, an approach often advocated by economists, may not be an effective substitute for our present arrangements, at least in the face of low labor demand. When given the resources and freedom to choose their own training strategy, low-income breadwinners appear to be no better at selecting a winning strategy than are the administrators and training specialists who now run training and employment programs.

THE NATIONAL SUPPORTED WORK DEMONSTRATION

The commitment made in the 1970s to assist hard-to-employ workers in finding jobs is perhaps best illustrated by the Supported Work program. It was a research and demonstration project, rather than a comprehensive employment program. It began in 1975 and was scheduled to last five years from its inception. Its basic objective was to provide individuals who had severe employment problems with work experience of about one year. The experience was provided under conditions of gradually increasing demands, close supervision, and work in association with a crew of peers. The guiding principle of the demonstration was that "by participating in the program, a significant number of people who are severely handicapped for employment may be able to join the labor force

and do productive work, cease engaging in socially destructive or dependent behavior, and become self-supporting members of society."¹²

Four groups of employment-handicapped workers were eligible for the program: women who were long-term recipients of AFDC, convicts recently released from prison, former drug addicts, and young school dropouts who often had a delinquency record. Fifteen sites were chosen for the program; ten were utilized for research on program effects. While each site was given responsibility for defining the type of work on which it would focus and the source of local funds on which it would draw, the entire program had a common research-evaluation emphasis. Hence, a variety of factors were standardized across the sites. These included the basic program design of low supervisor-participant ratios; steadily increasing standards of attendance, punctuality, and productivity; crew work and peer group support; and common eligibility criteria, wage rates, and employment duration. Like the Seattle-Denver experiment, the Supported Work Demonstration used a rigorous experimental design involving the random assignment of applicants to experimental (participant) and control (nonparticipant comparison) groups. We can therefore place substantial confidence in the demonstration's findings.

Over its five-year life, the demonstration provided services to over 10,000 persons, although at any one time the number of participants at a site was limited to 300. The evaluation of the demonstration is based on interviews with 3,214 participants and 3,402 controls. Each person in the research sample was interviewed prior to participation and given up to four additional interviews at nine-month intervals.

The participants suffered severe employment handicaps. Fewer than one-third had graduated from high school, most were black or Hispanic,

fewer than one-quarter were married, the number of weeks worked in the year prior to enrollment averaged six or seven, and (except for the female welfare group) arrest rates ranged from 54 to 100 percent. The work that was provided varied across sites, but included home rehabilitation, recapping tires, building furniture, and operating day care centers. Some program outputs were sold in the market in order to raise revenues for the program.

The program performance of the four enrolled groups varied considerably. Supported Work proved most effective in preparing the welfare women, who had the least work experience, for gainful employment. It also had a significant impact on the ex-addict group. For the ex-offender group, the results were marginal and not statistically significant, and no long-term positive results were found for the group of young dropouts. Overall, the participants in the program stayed an average of 6.7 months, even though the goal of the demonstration was about 12 months of participation. Thirty percent of the participants were fired because of poor performance; an equivalent number, however, moved on to full-time regular jobs. (The successful transition rate improved steadily over the course of the program.) About 10 percent (25 percent of the long-term welfare women) remained in the program for 12 months (in some cases 18 months), the maximum permissible program stay. These workers then had to be released from the program. The average cost to the public per recipient was \$5,740, but because most participants stayed in the program less than one year, the average cost per service year was over \$10,000. This cost declined steadily over the five years of the demonstration and is about the same as the service-year cost in another targeted training program, Job Corps.

The program had a variety of impacts on its participants. The AFDC group showed the most consistently positive response to the demonstration. The participation of these women was associated with increases in employment rate, hours worked, and earnings, both during and after the program period. In addition, there was a significant reduction in welfare dependency as well as reductions in the average amount of food stamps and other transfers received. The welfare women helped most by Supported Work tended to be older (between 36 and 44), to be less educated, to have been on welfare for a longer period, and to have little or no prior work experience. At least the last three of these effects would have been difficult to predict prior to the program, and indeed are somewhat surprising.

Among ex-addicts the demonstration raised employment and reduced criminal activity, but failed to have a statistically significant impact on drug use. The main impact on criminal behavior seems to have been concentrated in the first 18 months after enrollment in the demonstration. The program's effect on employment probably persisted for longer than that.

Ex-convicts in the demonstration do not seem to have been helped as much as the two groups just mentioned. The demonstration did not affect employment, welfare dependence, drug use, or criminal activity after participation ended. Similarly, the youth enrollees were not helped much, if at all, by the program. In this case, however, the evaluators found evidence that the target group was probably more employable than originally believed. At some time during the period of the study, between 80 and 90 percent of youth dropouts in the control group held a job. This

level far exceeds the rate of the other three control groups studied, indicating that the youth group was less disadvantaged than the other target groups.

The Manpower Demonstration Reserach Corporation and Mathematica Policy Research conducted a very careful benefit-cost evaluation of the demonstration. They computed the benefits and costs of the program from three different perspectives--that of program participants, that of taxpayers, and that of society as a whole (participants and taxpayers). The social benefits include the output produced by workers in the program, increases in their post-program earnings, reductions in criminal activities, and savings from reduced participation in other public employment, training, or drug treatment programs. The social costs include all program operating costs (excluding transfer payments, however, because these are simply a redistribution of income). The benefit-cost tabulations were based on extrapolations over the typical working life of the participants, with benefits assumed to decay at a rate of 50 percent every five years, except among AFDC mothers, where no decay rate in benefits was assumed.

The benefit-cost analysis showed that the demonstration had considerable net social payoff for the welfare mothers enrolled, primarily due to the long-term earnings gains assumed and the value of the output from the demonstration jobs. Benefits also exceeded costs for the ex-addicts, in large part because of the reduction in socially destructive behavior (i.e., crime) and the gains in employment and earnings. For ex-convicts the results were less conclusive. The net benefit of the

program may have been positive or negative depending on the assumptions used to value the benefits of the program. Not surprisingly, in view of the estimated impact of the demonstration on youths, the program's cost was found to outweigh its benefits for the youth dropout group.

Because of the very specific nature of the treatment that was tested in the Supported Work demonstration, it is difficult to draw broad policy conclusions from its results. The finding that its approach had the greatest payoff in the case of AFDC mothers is consistent with a few other findings from the last decade of research on training and employment programs. Some of the studies of the Continuous Longitudinal Manpower Survey (CLMS) have also concluded that disadvantaged women helped by the Comprehensive Education and Training Act (CETA) appeared to have obtained the greatest program benefit. Similarly, in the Seattle-Denver experiment, the only group to show a positive impact from the counseling program (as distinct from the voucher program) was the sample of single women with children. Also, as we shall see below, the Employment Opportunity Pilot Project appeared to have a more consistently and significantly positive effect on unmarried women than on other groups served. It would thus appear that single mothers are more susceptible to being helped by public training and employment efforts than other groups of hard-to-employ workers.

THE EMPLOYMENT OPPORTUNITY PILOT PROJECT

The history of the Employment Opportunity Pilot Project--or EOPP--was a tumultuous one, marked by shifting objectives and premature cancellation. It is said that we learn from our mistakes. If this were

true, EOPP should have been one of the most richly informative demonstrations ever undertaken.

In substance it was a demonstration, and not an experiment, in contrast with the two efforts discussed above. The project was begun by the Carter administration in order to estimate participation rates and potential effects of a guaranteed jobs program similar to that proposed in Carter's welfare reform package. Alarmed by the work-effort reductions estimated in the Seattle-Denver experiment, the administration was determined to limit the work disincentive effects of its welfare proposal by requiring certain welfare recipients to accept public service employment (PSE) if they were unable to obtain unsubsidized private sector jobs. The president's welfare reform efforts were twice rebuffed by Congress, but his PSE proposals were treated more sympathetically. In 1978 Congress permitted the Department of Labor to set up a fourteen-site pilot test of a guaranteed jobs program.

Even before the first EOPP enrollments took place in 1979, the basic objectives of the demonstration had already been modified. This was in part due to the administration's evolving objectives in reforming welfare and CETA. In addition to simply providing a test of the guaranteed jobs concept, which was expected to be very expensive, the architects of EOPP also hoped to test new approaches to job finding among the hard-core unemployed. If applicants for PSE jobs could be required to participate in intensive and structured programs of job finding, and if those programs turned out to be successful, the "demand" for PSE job slots, and hence the cost of PSE, could be limited.

At the time the demonstration began in 1979, its objective was to determine whether a program that provided a combination of

job search assistance and subsidized employment and training could succeed in increasing the employment and, hence, reducing the welfare dependence of adults in low-income families with children. The program, targeted primarily toward families that were receiving AFDC, provided participants with intensive job search assistance and support services, such as child care and transportation assistance. Participants who were unsuccessful at finding an unsubsidized job after a prescribed period of active search were offered a subsidized job or training.¹³

When President Reagan took office in 1981, the goals of the program, or at least the focus of the program evaluation, shifted once again. The new administration wished to abolish public service jobs, not to pilot test a program that guaranteed them. It emphatically signaled this goal first by ending enrollments into EOPP's PSE program and sharply curtailing enrollment in other components of EOPP, then by prematurely terminating the entire project in October 1981, less than two and a half years after operations began in 1979. Mathematica, the prime research contractor for the project, was directed to discover the impact, if any, of EOPP's job search assistance program and to provide a cost-benefit analysis of that program.

The implementation of EOPP and its evaluation were seriously harmed by these shifts in program objectives. The original research and implementation design was sensible for a pilot test of a guaranteed jobs program. It was extremely deficient, however, for evaluating alternative approaches to job search assistance, the goal emphasized in the final evaluation contract. The available control group was ill-suited for this purpose.

To evaluate a guaranteed jobs program, the original goal, it is necessary to conduct saturation demonstrations under a variety of local labor market conditions. By "saturation demonstrations" we mean that the program must be offered on an unlimited basis to all income-eligible families in a particular community. Saturation is required in order to determine participation rates in a well-publicized program and, equally important, to see whether such a program may seriously disrupt local labor markets by driving down the available supply of labor for unsubsidized employment. To see how local labor market conditions were affected by EOPP, it was necessary to obtain a basis for comparison. Mathematica and Department of Labor officials selected fourteen comparison sites to be used as a "control group" for the fourteen pilot sites in the demonstration. (Because "control sites" were selected, EOPP might arguably be called an experiment rather than a demonstration project. However, eligibility for treatment was not randomly assigned to individuals except in Dayton and Philadelphia, and hence the project was probably closer to an ordinary demonstration than to a formal social experiment.) This strategy required massive amounts of household interviewing in both pilot and comparison sites.

Only a small proportion of these household interviews would have been needed for an adequate assessment of the job search assistance program by itself. Indeed, to test that program, an experimental design involving at most a few thousand individuals in selected labor market environments, randomly assigned to program and control groups, is all that would have been required. Neither saturation, nor multiple control sites, nor massive interviewing would have been necessary.

EOPP was administered by the state and local officials (prime sponsors) responsible for administering local CETA programs. The competence and commitment of local administrators thus varied considerably. The prime sponsors were responsible for publicizing the availability of EOPP services, identifying, recruiting, and determining the technical eligibility of potential clients, providing support services like child care for enrolled participants, establishing and administering a structured program of job search assistance, and providing public service jobs, work experience slots, and classroom and on-the-job training (OJT) opportunities for clients unable to obtain unsubsidized employment. The broad character of program responsibilities and the potential for administrative discretion at each point are noteworthy, and they threaten the reliability of evaluation findings. We simply cannot be confident about the exact nature of the treatment as delivered in the field.

EOPP tested self-directed job search methods that are quite distinct from the job referral and job development techniques usually used in the U.S. Employment Service or CETA. Clients were taught effective methods of job search and encouraged to follow a rigorous and structured routine in looking for employment. People who could not find unsubsidized jobs in five to eight weeks were offered a subsidized employment or training position, which could last up to one year before workers or trainees were recycled through the job search assistance program. Workers in PSE jobs and OJT positions were paid regular wages, while those in work experience or classroom training slots were given a weekly training stipend.

To be eligible for EOPP job search assistance, applicants had to be adult members of families that included one or more children and that

either received AFDC or had income below 70 percent of the Lower Living Standard developed by the Bureau of Labor Statistics. To be eligible for subsidized employment or training, individuals were required to complete the job search phase of the program without obtaining unsubsidized employment and, in addition, to be the family's primary earner and either receive AFDC or have low enough income to qualify for AFDC. In most sites the program was aimed primarily at adult AFDC recipients.

Mathematica's evaluation of EOPP covers only ten of the fourteen communities involved in the demonstration. In those ten communities it estimated that over 190,000 adults were eligible for EOPP services at some point during the demonstration.¹⁴ However, of that total only 120,000 were eligible for the full range of EOPP services, including subsidized employment and training. Only 21,000--18 percent--of those fully eligible chose to enroll. An additional 2,000 adults eligible only for job search assistance also enrolled in the program.¹⁵ Of those individuals who filled out the enrollment forms, only about 62 percent remained in the program long enough to receive some job search assistance. One-third of the people receiving that help obtained an unsubsidized job. Only 4,100--17 percent of enrollees--remained with the program long enough to receive subsidized employment or training, of whom approximately two-thirds were assigned to PSE jobs.¹⁶ Thus, of the 120,000 potential participants in EOPP's "guaranteed jobs" program, fewer than 3 percent actually obtained PSE jobs.

The striking feature of these statistics is the very small proportion of program eligibles who actually received program services, especially very expensive services like subsidized jobs and training. This suggests

that a guaranteed public jobs program aimed at the welfare-eligible poor would be considerably less expensive than anticipated by the Carter administration, which expected a much higher participation rate. On the other hand, the program would also be much less successful than expected in reducing welfare dependence, since only a small percentage of AFDC recipients would apparently be forced to participate in such a program.¹⁷ In part the low participation rate in the jobs program was attributable to uncertain guidelines from the Labor Department, poor program administration at the local level, normal start-up problems, and a lack of publicity for the program. Even with these problems, it was astonishing to program operators that so small a proportion of obviously eligible people chose to enroll. Among AFDC recipients who were mandatory participants in the Work Incentive (WIN) program (and thus likely to be ready to hold a job), only one-third enrolled in EOPP, and yet the availability of EOPP was widely advertised among that group.¹⁸ Among nonrecipients of AFDC who were eligible for the PSE jobs, only 8 percent enrolled in the EOPP program.¹⁹

In view of the apparently generous offer provided by the program, this indifference to EOPP is interesting. Of course, it is possible to keep enthusiasm for public jobs down by erecting enough bureaucratic hurdles--a complex and lengthy application process, mandatory participation in a job search program, and potentially long delays before assignment to a PSE job. Nonetheless, it appears that the attractiveness of a temporary PSE job paying between one and two times the minimum wage is not nearly as great as sometimes assumed. Even though EOPP provided a

highly imperfect test, the administration of the demonstration was probably not perceptibly inferior to what would be provided in an ongoing program. The local administrators of the program were, after all, the same people responsible for administering CETA and are probably now running training and referral programs under JTPA. If there is any future consideration of a guaranteed jobs program for welfare recipients, EOPP has taught us that both the costs and the benefits will be considerably below what was expected in the mid-1970s.

What of the other objectives of the project? The evaluation contractor concluded that the job search assistance program run by EOPP was probably effective in helping participants find jobs. Enrollees in that program increased the amount and effectiveness of their search efforts. In comparison to unemployed workers in the target population who did not enroll in EOPP, participants spent nearly twice as many hours a week searching for a job, contacted about four times as many potential employers, and filed approximately 75 percent more formal job applications.²⁰ As mentioned earlier, about one-third of enrollees receiving job search help landed an unsubsidized job. Although it is unclear how much of an improvement is indicated by this placement rate, Mathematica concluded that for the largest group of enrollees--single mothers--EOPP probably raised the employment rate by 10 to 12 percentage points and raised the probability of unsubsidized employment by 7 to 9 percentage points.²¹

Because EOPP was so poorly designed to measure the effectiveness of job search assistance, Mathematica could not determine the fraction of the employment gain that was due solely to the job search plans tested.

Nor were the researchers able reliably to measure the impact of EOPP on the other groups served--married women and men with dependent children. Mathematica could detect no impact of the program on welfare dependence, a surprising finding in view of the population served by EOPP, which consisted overwhelmingly of public assistance recipients. Because EOPP and its evaluation were terminated with unseemly haste in 1981, we will never know whether the employment gains registered by EOPP participants were temporary or long-lasting. Nor can we ascertain whether welfare dependency was eventually affected by the program. Because of the limitations described above, Mathematica was unable to perform a benefit-cost analysis of the job search program alone, although the analysts did conclude that the project's overall social benefits probably exceeded its social costs. Based on our reading of the evidence, it appears that a modest and comparatively inexpensive program to help low-income breadwinners search for work may reduce spells of unemployment and raise the fraction of time spent working. Even though it is doubtful that this kind of help will change many workers' lives or radically change the nature of jobs they obtain, the help is nonetheless worthwhile, and it comes at relatively low cost.

Before concluding this discussion of EOPP, we should also note that some of the pilot sites tested variants of a basic self-directed job search model. One of the most interesting variants was tested in Dayton, Ohio, where wage-subsidy vouchers were distributed to a randomly selected subgroup of enrollees in the job search classes. The vouchers were simply certificates provided to participants to help them in their search for work. Participants were encouraged to alert potential employers of

their vouchered status. If a vouchered job seeker was hired by a qualified employer, the employer could claim a subsidy for a fraction of the wages paid to the newly hired worker. The subsidy was payable either in the form of a tax credit or a direct check payment to the employer. It was worth up to \$4,500 over a two-year period.

In effect, the vouchered workers were "on sale." Employers, however, appeared to regard these workers as damaged goods. In comparison to unvouchered participants in the EOPP program, vouchered job seekers were significantly less likely to obtain employment during their five- or eight-week job search period. Although this "experiment" is limited in many ways, and the research on it was discontinued too early to be definitive, the findings are intriguing. The basic result appears to show that a targeted wage voucher may hurt rather than help a job seeker's chances of employment. Perhaps it should come as no surprise that our nation's two most important wage subsidy programs--WIN and Targeted Jobs Tax Credit--are so little used. Because the stigma associated with these programs may outweigh their tax advantages to employers, the unemployed may be reluctant to use them and employers may be less likely to hire job seekers who offer them.

A MORAL AND SOME LESSONS

We now turn from the specific findings of these three projects to some of the general lessons that can be drawn from them. Social experiments have primarily been tools of social scientists seeking guidance for effective policy reform or innovation, but their conclusions have often been very pessimistic for those wishing to change public policy.

According to the Foreword of the volume that summarized results of the New Jersey Income Maintenance Experiment, the decision to undertake that experiment was based on the "rapid spread of the belief, especially among economists, that negative income taxation was an idea whose time had come."²² After the New Jersey Experiment began, two presidents, Nixon and Carter, proposed variants of a federal negative income tax, but in neither case was the cause of the proposal advanced by findings from the experiments. In fact, the high price tag pinned to the proposed Carter plan, which certainly harmed its chances of enactment, was estimated using interim results from the Seattle-Denver experiment.

Because of the rigor with which experiments are designed and evaluated, there may be a bias toward reaching pessimistic conclusions about policies that are experimentally tested. The tested program is subject to critical examination of a type that is rarely imposed on existing programs. Such an examination is likely to reveal undesirable or even pernicious side-effects of a policy that might not otherwise be detected. Consider, for example, the Earned Income Tax Credit. Under this apparently benign provision of the tax code, tax credits are provided workers with low annual incomes who have dependents. The purpose of the credit is to encourage work effort. If this policy were systematically evaluated, using the methods applied to social experiments, the credit might be shown to reduce work effort or encourage family dissolution, as the NIT was found to do. Indeed, the credit may yield a net increase in work disincentives because it increases marginal tax rates for more workers than are eligible for a subsidy on marginal work.²³ If these effects were found to occur, and if they were widely publicized, the

credit could be politically doomed. However, such effects are unlikely to be investigated because of the program's uncontroversial nature.

Numerous other examples could be mentioned. Do subsidized student loans stimulate increases in years of education? If they do, is the added investment in education worth its social and private cost? Do business tax reductions and other state and local subsidy programs to attract new businesses achieve their goals? Such programs could conceivably reduce or delay local investment projects if businesses delayed their decisions as a result of their efforts to attract subsidy support.

If an experimentally tested program fails to achieve its intended purpose, or if it has disagreeable consequences, those facts can be demonstrated with statistical rigor. Even more disturbing, if the program fails to achieve spectacular positive results, the degree to which it falls short of perfection can be measured precisely and then used as an argument against its implementation. If, on the other hand, an ongoing program does not achieve its objectives or does harm, its failure may remain unsuspected, or at least unproved.

As an empirical fact, policies about which there is strong disagreement are the ones most likely to be subject to rigorous experimentation--negative income taxation, housing vouchers for the poor, national health insurance, and labor market assistance to low-income workers. Programs aiding the able-bodied poor are among those with the weakest popular mandate, and hence their reform will nearly always inspire deep controversy. It is unclear whether experimentation per se can shed much light on the main points at issue--the demands of equity, the nature of a fair distribution, and the limit of society's obligation

to help those who are at least partly able to help themselves. Our experience in the last fifteen years has taught us that large-scale social experiments can be relied on to teach us something of value about the policy in question, but what we are taught can seldom be relied on to aid the cause of reforming or improving policy. Society is not even-handed in subjecting programs for the poor and nonpoor to experimental investigation. It has not examined transfers to the nonpoor with the same degree of intensity as it has examined those to the poor. We should therefore not be surprised that experimental scrutiny has been less kind to programs designed to benefit the poor. There is a moral here, and it is illustrated in the three experiments we have considered: if you advocate a particular policy reform or innovation, do not press to have it tested.

Beyond this political and economic moral, are there lessons for research or evaluation that can be gleaned from the experiments? One such lesson concerns the costs and benefits of large-scale social experimentation relative to nonexperimental social research. Clearly, the research costs of social experimentation are enormous. For the three experiments reviewed here, a conservative estimate of the costs of program administration (including experimental transfers, stipends, and wages) and evaluation would exceed \$200 million. The potential benefits in terms of additions to knowledge may also be substantial, especially when it is recognized that obtaining reliable information about human behavior is usually a slow process. However, if the opportunity cost of any proposed experiment is a reduction in nonexperimental research costing the same amount of money, the expected findings would have to be

extremely valuable for the benefits of an experiment to exceed its cost. Of course, this conclusion is weaker if the opportunity cost of the resources used for experimentation is low. This would be the case, for example, for resources that are diverted from some activity with low social value.

In view of the high cost of experiments, it is appropriate to subject proposals for future experiments to a test that includes the following questions:

1. Have adequate models of the behavior which the experimental treatment is designed to affect been developed and tested on existing bodies of data?
2. Can the experiment and its evaluation meet high standards of basic research? That is, can problems of time horizon, contamination, replicability, and extrapolation of results to a national program be handled adequately in the experimental design or in the evaluation of experimental results?
3. Can the experiment provide evidence about a social policy that cannot be obtained using less expensive, non-experimental methods? Alternatively, can the experiment provide findings that are sufficiently more reliable or statistically precise to justify the added cost of the research?
4. How important are the potential research findings about experimental outcomes? Are they crucial in determining whether the tested treatment is a good or bad policy?
5. Can the experiment permit tests and evaluation of the operational feasibility of social policy measures and yield evidence on the effectiveness of alternative administrative arrangements of such programs?
6. Can the experimental findings be validly generalized to infer the consequences of policies not specifically tested in the experiment?

The number of potential social experiments that can pass the test implied by these questions is not likely to be large. This conclusion is

strengthened by our review of the findings of the three experiments. While the evidence on behavioral responses is more reliable than is likely to be obtained from nonexperimental research, its value, in terms of added knowledge per dollar of cost, is not unusually high except in the case of the tested NIT plans. For the training and employment experiments, including the one run as part of the Seattle-Denver experiment, the programs tested were so specific in nature that it is difficult to extrapolate the findings to any other programs except those that are run exactly as they were. (For EOPP even this may be impossible because the tested treatments are essentially nonreplicable.)

The Seattle-Denver NIT experiment has proved more valuable for two reasons. Its findings are considerably more reliable and statistically precise than any that had been obtained in the preceding ten years of nonexperimental research. Moreover, its findings are useful in evaluating tax and welfare policies in addition to those actually tested in Seattle and Denver, in part because there is a well-developed theory for assessing labor supply responses to tax rates and guarantees.

But the exception represented by the Seattle-Denver experiment is rare. Many conceivable experiments in the field of employment and training must concentrate on testing "black box" treatments. In a black box experiment, a treatment is provided and an outcome observed, but the process by which the treatment affects the outcome is largely unknown. Supported Work and the job search model tested in EOPP both represent this kind of treatment. There is no well-established theory, as existed in the case of the NIT experiments, that permits us to predict whether and how these particular approaches will affect participants. Nor can we

predict from experimental findings the effects of similar--but not identical--policy options. This lack of knowledge regarding the process by which treatment affects performance limits the applicability of the findings. In the case of both Supported Work and EOPP, the treatment tested was of little interest by the time the research was completed, and the findings, in turn, were of limited value in assessing policy options then being considered.

Black box experiments can be valuable in employment and training research if they are relatively inexpensive but rigorous and if there is systematic variation in the treatments which are tested. Investing large sums of money to test a single approach is likely to be a serious error except under very unusual conditions. To justify its high cost, a social experiment must offer the prospect of valuable additions to knowledge about human behavior. In light of the moral mentioned above, the benefits of an experiment will seldom include basic reforms to policy.

NOTES

¹A very small number of families were enrolled into experimental plans lasting twenty years.

²U.S. Department of Health and Human Services, Office of Income Security Policy (Overview of the Seattle-Denver Income Maintenance Experiment Final Report (Washington, D.C.: Government Printing Office, 1983), pp. 13-16. The higher estimate of the impact on single women heading families is based on the responses of women in the five-year group during the fourth and fifth experimental years. Remaining estimates are based on reported responses of enrollees in both the three-year and five-year groups during the second and third experimental years.

³Ibid., p. 6.

⁴See Henry Aaron and John Todd, "The Use of Income Maintenance Experiment Findings in Public Policy, 1977-78," Proceedings of the Industrial Relations Research Association Annual Meeting (Madison, Wis.: IRRA, 1979), pp. 46-56.

⁵Implementing an NIT program for single-parent families, given the combination of existing transfer programs, is difficult. Because of the widely varying AFDC benefit levels across states, it is difficult to select an NIT guarantee level that is low enough to be affordable, but high enough so that only a small fraction of families in the high-benefit states receive an NIT payment that is no lower than their current benefit. A national NIT plan with a guarantee equal to three-quarters of the poverty line would increase labor supply among single mothers, not because of the work incentive embodied in a low tax rate, but because

transfer benefits would be slashed for so many mothers in states currently paying high benefits.

⁶This result and those reported in the preceding paragraph are from Philip K. Robins and Richard W. West, "Labor Supply Response," in U.S. Department of Health and Human Services, Final Report of the Seattle-Denver Income Maintenance Experiment, Vol. 1, Design and Results (Washington, D.C.: Government Printing Office, 1983), pp. 180-187.

⁷Strictly speaking, the experiment provided no evidence about the impact of raising marginal tax rates to 100 percent. Within the range of tax rates tested in the experiment, however, higher tax rates appear to be associated with higher aggregate labor supply (*ibid.*, p. 182).

⁸Assignments to the schooling and employment-training programs were conducted in such a way that analysts were able reliably to distinguish the separate impacts of those programs and the tested NIT plans.

⁹Reimbursable (or direct) expenses included costs for tuition, books, transportation, and child care.

¹⁰Note that this was the impact on program eligibles; the impact on program participants was of course much greater. The 50 percent subsidy also encouraged some extra schooling, but the increases were smaller. See Bureau of Social Science Research, "Vouchering Manpower Services: Past Experiences and Their Implications for Future Programs," report to the National Commission for Employment Policy, Washington, D.C. (1982), p. 20.

¹¹*Ibid.*, p. 29.

¹²This quote (p. 1) as well as much of the material for this section is drawn from Manpower Demonstration Research Corporation, Summary and

Findings of the National Supported Work Demonstration (Cambridge, Mass.: Ballinger Publishing Co., 1980).

¹³Mathematica Policy Research, Inc., Final Report: Employment Opportunity Pilot Project: Analysis of Program Impacts (Princeton, N.J.: MPR, 1983), p. 1.

¹⁴Ibid., p. 20.

¹⁵Ibid., p. 22.

¹⁶Ibid., pp. 27, 105, and 116.

¹⁷We should emphasize that the low participation of welfare recipients in the demonstration was partly attributable to poor enforcement of job search requirements in local welfare departments. If the job search/PSE jobs and welfare programs had been more tightly coordinated, the costs and hence potential benefits of an EOPP-type program might have been greater.

¹⁸Mathematica Policy Research, Final Report, p. 22. Many mandatory participants in WIN are in fact required to participate in an activity like EOPP as a condition for continued receipt of welfare benefits.

¹⁹Ibid., p. 22.

²⁰Ibid., p. 108.

²¹Ibid., p. 3. A small percentage of enrollees obtained employment in EOPP's own jobs program. For that reason the gains in unsubsidized employment were smaller than those in all forms of employment.

²²David Kershaw and Jerilyn Fair, The New Jersey Income Maintenance Experiment, Vol. I, Operations, Surveys, and Administration (New York: Academic Press, 1976), p. xi.

²³This occurs because the 10 percent work incentive subsidy rate applies to annual earnings from 0 to \$5,000, an earnings range where the population density is thin, while the benefit reduction rate of 12.5 percent, which is a work disincentive, applied to annual earnings from \$6,000 to \$10,000, where the population density is thicker.