The Role of Randomized Field Trials in Social Science Research

A Perspective From Evaluations of Reforms of Social Welfare Programs

ROBERT A. MOFFITT
Johns Hopkins University

One of the areas of policy research where randomized field trials have been utilized most intensively is welfare reform. Starting in the late 1960s with experimental tests of a negative income tax and continuing through current experimental tests of recent welfare reforms, randomized evaluations have played a strong and increasing role in informing policy. This article reviews the record of these experiments and assesses the implications of that record for the use of randomization. The review demonstrates that the usefulness of randomized field trials in the area of welfare reform has been limited by a number of weaknesses, some of which are inherent in the method and some of which result from constraints imposed by the political process. The conclusion is that randomized field trials have an important but limited role to play in future welfare reform evaluations and that it is essential that they be supplemented by nonexperimental research.

Keywords: *social experimentation; welfare programs; program evaluation; poverty*

Unlike the case in many other social sciences, randomized field trials (RFTs) have been used extensively in certain subareas of the discipline of economics. Although there are several such subareas where experimentation has been employed, the area of social welfare is perhaps that which has seen the most intensive use. RFTs in social welfare were begun in the 1960s with experimental tests of a negative income tax, and RFTs testing various reforms of cash welfare—most notably, reforms to the Aid to Families with Dependent Children (AFDC) program—have continued unabated since then and have, indeed, accelerated in the 1990s. RFTs have been extended to the estimation of effects of

Author's Note: *This article is a revised version of an article prepared for the Conference on Randomized Experimentation in the Social Sciences, Yale Institution for Social and Policy Studies, New Haven, August 20, 2002. The author would like to thank Gordon Berlin, Donald Green, David Greenberg, Alan Krueger, Charles Michalopoulos, and Howard Rolston for comments. All views are those of the author alone.*

AMERICAN BEHAVIORAL SCIENTIST, Vol. 47 No. 5, January 2004 506-540
DOI: 10.1177/0002764203259292

programs for job training, housing, health insurance, Food Stamps, Medicaid, unemployment insurance, and earnings and wage subsidies. Millions of taxpayer dollars have been spent on these experiments, representing a major social investment in knowledge accumulation. Given this long and rich history, it seems fitting to assess the contribution of RFTs in this area as a demonstration of how useful the methodology can be and to draw lessons that might be a partial forecast of how much experimentation might be able to contribute in other areas. In addition, a review of a specific area of RFT research, such as that presented here, may help make progress in the experimental-nonexperimental debate, which generalized, abstract discussions cannot.

A review of all the RFTs in the area of social welfare is far beyond the scope of this article. Instead, the review concentrates on RFTs that have tested reforms in the cash welfare program AFDC. Even here, there are too many to completely enumerate and only the most important, and most influential, RFTs are reviewed. Still, given the volume, importance, and influence of experiments in this corner of social welfare, the review is still capable of demonstrating some general lessons.

The article is composed of two main sections. The first summarizes, albeit briefly, the most important experiments on the AFDC program throughout the past 30 years. The second section provides a discussion of the strengths and weaknesses of the experiments and draws lessons for the experimental methodology.

A BRIEF REVIEW OF RFT IN THE AREA OF CASH WELFARE

Welfare reform. A thumbnail history of welfare reform developments in the AFDC program throughout the past 40 years is useful for those completely unfamiliar with the area.[1] The AFDC program was created by Congress in 1935 as part of the Social Security Act. The program provided cash benefits to low-income families where children were present and where one biological parent was absent from the household. The main group intended for support were poor widows with children and, indeed, the major recipient group for the first 20 years of the program were such single-mother families. The program experienced little reform over that period and caseloads grew more or less in line with population. However, beginning in the late 1950s and early 1960s, the proportion of the caseload composed of single mothers who were divorced or separated began to grow and, simultaneously, the political popularity of the program began to decline. Taxpayers and their elected representatives became interested in reforming the program in a variety of ways, many of which were intended to increase work levels of recipients. This goal was reinforced by a rising labor force participation rate of women as a whole that occurred at the same time, because this created a presumption that women should not necessarily remain in

the home to raise children—the original rationale for support for widows—but should, if means were available, participate to at least some degree in the workforce.

These pressures led to a series of work-related reforms. In the late 1960s and early 1970s, reforms generally took the form of providing financial incentives by lowering the benefit-reduction rate, or tax rate, in the AFDC program. This was a reform suggested by Milton Friedman in the 1960s as part of his formulation of a negative income tax (Friedman, 1962) and later adopted and promoted by many other economists, including Robert Lampman and James Tobin (Lampman, 1965, 1968; Tobin, 1966; Tobin, Pechman, & Mieszkowski, 1967). These financial incentives also went by the name of "enhanced earnings disregards" because the way that financial incentives were provided was by disregarding some of the earnings that welfare recipients obtained when calculating benefits. In the AFDC program at that time, the tax rate was 100%, implying that a recipient who earned an extra $100 would have her benefit reduced by exactly $100, leaving her overall income unchanged and hence providing no incentives to work. The negative income tax proposed a tax rate less than 100%; for example, if it were 50%, an extra $100 in earnings would result in only a $50 reduction in benefits because $50 of the earnings would be disregarded, leaving overall income $50 higher and consequently giving the recipient some reward for working more. A negative income tax program was proposed by the Nixon Administration and passed the House of Representatives but failed in the Senate; this proved to be the political high water mark of the idea. As subsequent Presidential administrations in the 1970s considered the negative income tax and alternatives, interest shifted to policies that induced work by means of work requirements rather than financial incentives. States showed great interest in work requirements in the 1980s, testing a variety of types of such policies, and in 1988, Congress passed legislation that mandated that certain fractions of recipients be involved in some kind of work, training, or education activity, with considerable emphasis on training and education.[2]

However, because caseloads continued to grow after 1988 and because work levels among AFDC recipients remained low even after this legislation, policy took a very different direction in the early 1990s, shifting toward much stronger work requirements backed up by sanctions (i.e., full or partial benefit reductions for noncompliance), toward fewer exemptions from work requirements, and toward the imposition of time limits on the length of benefit receipt.[3] Time limits are, in a sense, a final answer to the question of how to promote work and decrease welfare dependence by literally making families ineligible for benefits after a given length of time. In the early 1990s, virtually every state began adopting these reforms, and by 1996, the majority of the nation's AFDC caseload was already subject to some type of new program with these elements. Congress took action in 1996, passing the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA), which converted many of these reforms into federal law and, hence, imposed them nationwide. The Act mandated federal

time limits, minimum work requirements, the imposition of sanctions for noncompliance, and converted the program to a block grant, devolving much responsibility for program operation and design to the states. The Act also abolished the AFDC program and replaced it with the Temporary Assistance for Needy Families (TANF) program.

Randomized field trials. Tables 1 to 4 provide a selective review of the RFTs throughout this period, which tested reforms of the AFDC program or related cash welfare reforms.[4] Table 1 lists the design features of the major RFTs in the 1960s, 1970s, and 1980s and Table 2 summarizes their results. The first four listed are the negative income tax (NIT) experiments, which began in the late 1960s and continued into the 1970s (see Burtless, 1987; Moffitt & Kehrer, 1981; SRI International, 1983, for reviews). The NIT experiments were in many respects unique, and perhaps the most ambitious of all the RFTs in this area ever conducted. Operating outside the regular AFDC system, and testing a proposed welfare program that had emerged among academics and therefore was fairly abstract in concept, the experiments tested benefit schedules with different welfare guarantees (the amount paid to a family with no income or earnings) and different tax rates. The control group in the experiments received the existing AFDC program, for single mothers, or no program at all, for married men and women (who were covered only minimally by the program). The object was to determine if reduced tax rates increased work levels, as presumed by advocates of an NIT. The results showed, perhaps surprisingly, that lowered tax rates had essentially no effect on labor supply, a result that should have been anticipated by economic theory but was not.[5] However, in the political process, Congress and the public ended up focusing most of its attention on the effect of an NIT on work levels relative to no program at all, and here the experiments showed that an NIT—similar to any welfare program without work requirements—would reduce work effort, as predicted by economic theory.[6] The failure of the Nixon Administration's NIT proposal in Congress has been partly attributed to Congressional realization that the existing AFDC program had work disincentives and its consequent disinterest in any new program that had the possibility of merely reducing their size (Moynihan, 1973).

From a design standpoint, the NIT experiments had many critics (e.g., Ashenfelter & Plant, 1990; Hausman & Wise, 1985b; Moffitt & Kehrer, 1981; Pechman & Timpane, 1975). The experiments were criticized for inadequacies in technical allocation design, inability to address biasing attrition and underreporting, and for the econometric methods used to analyze the data. The technical allocation design, for example, assigned individuals to the control group and to the experimental group partly on the basis of their preexperimental income levels and randomized only within income strata, but with experimentals and controls distributed unequally across strata. This meant that it was impossible to analyze experimental-control differences by simple raw differences in means between the groups because income had to be controlled for

statistically before doing so. A number of the experiments also attempted to test too many alternatives (e.g., 48 different cells in the Seattle-Denver experiment), resulting in inadequate sample size and low statistical power in each. Attrition was also a severe problem in the experiments, particularly in New Jersey, where there was evidence that more low-income controls attrited than low-income experimentals, leading to a bias in experimental-control earnings and income differentials. Underreporting of income appeared to be a problem in several experiments as well, because low-income experimentals had more incentive to underreport work levels, earnings, and income to increase payments, and this would lead to a bias in the experimental-control differences in those variables. The analysis methods used in the experiments were criticized at various times for being either too structural or not structural enough; the New Jersey experiment was criticized for estimating only simple nonstructural models, whereas the Seattle-Denver experiment was criticized for insufficiently presenting simple mean differences and too often presenting the results of structural models (Ashenfelter & Plant, 1990).[7]

In retrospect, these weaknesses of the experiments were mostly the result of lack of experience by researchers in designing experiments and in the analysis of experimental data, because none is inherent in the experimental method. Indeed, all three of these issues have been directly addressed by subsequent experiments and have been largely eliminated as drawbacks.[8] It will be argued below that these weaknesses of the NIT experiments were not inherent ones and that because of other characteristics of those experiments to be described below, they nevertheless represent something of an ideal type that has never been achieved since.

A number of other RFTs followed the NIT experiments. Two major ones were the Supported Work Experiment and the AFDC Homemaker–Home-Health-Aide Demonstration, both of which were large-scale evaluations of expensive, innovative programs that broke in significant ways from past efforts at getting welfare recipients into the workforce. The Supported Work Experiment was an intensive attempt to nurture recipient work skills and to gradually introduce recipients into the stressful world of work, whereas the AFDC Homemaker–Home-Health-Aide Demonstration trained and put AFDC recipients to work as home health care aides. Both the Supported Work and AFDC Homemaker–Home-Health-Aide Demonstration are regarded in the literature as well-run experiments that yielded credible and interesting results. Both programs were successful in increasing earnings and reducing welfare recipiency, but both were too expensive for later policy makers to become enthusiastic about. But because they were large-scale and expensive, they were in this sense similar to the NIT.

Most of the other RFTs listed in Table 1, all of which were conducted in the 1980s, represented a major shift in design approach. These programs differed

(text continues on p. 520)

TABLE 1: Selected Randomized Field Trials of Reforms in Cash Welfare, 1960s to 1980s: Design

	Demonstration Period	*Description of Treatments*	*Population*
Negative income tax experiments			
New Jersey	1968-1972	Negative income tax plans with different guarantees and tax rates	Low-income households with one nondisabled man
Rural	1970-1972	Negative income tax plans with different guarantees and tax rates	Rural low-income households with a male head
Seattle-Denver	1970-1977	Negative income tax plans with different guarantees and tax rates, and one training treatment	Low-income married and female-headed households
Gary	1971-1974	Negative income tax plans with different guarantees and tax rates	Black families with at least one child younger than 18
National Supported Work Demonstration	1975-1978	Structured work program with graduated increase in work standards, peer support, and close supervision	Long-term AFDC recipients[a]
Maryland (Baltimore Options Program)	1982-1983	Package of job search, education, and work-related services	Newly mandatory AFDC recipients and new AFDC applicants
California (San Diego Job Search Demonstration)	1982-1983	Mandatory job search and work activities	New applicants to AFDC
Washington (CWEP)	1982-1983	Workfare and job clubs	AFDC recipients
Arkansas (Work Demonstration)	1983-1984	Mandatory job search and workfare	AFDC applicants and recipients
Maine (Training Opportunities in Private Sector)	1983-1985	Classroom training, subsidized work experience, on-the-job training	AFDC applicants with eligible characteristics

(continued)

TABLE 1 (continued)

	Demonstration Period	*Description of Treatments*	*Population*
Virginia (Employment Services Program)	1983-1984	Job search and work activities	AFDC recipients
West Virginia (CWEP)	1983-1984	Workfare	AFDC recipients
AFDC Homemaker–Home-Health-Aide Demonstration	1983-1986	AFDC recipients trained as home–health care aides for low-income elderly and disabled individuals	AFDC recipients
Florida (Trade Welfare for Work Program)	1984-1986	Mandatory job search, job search, and on-the-job training	AFDC recipients
New Jersey (Grant Diversion Project)	1984-1987	On-the-job training	AFDC recipients
Illinois (Cook County Job Search and Work Experience)	1985	Mandatory training and work activities	AFDC recipients
California (San Diego Saturation Work Initiative Model)	1985-1987	Job search, subsidized work experience, community education and training	AFDC recipients
Teenage Parent Demonstration	1987-1991	Case management, employment and training services	Teenage AFDC recipients

California (Greater Avenues for Independence)	1988-1990	Basic education, job search, and skills training	AFDC recipients
Washington (Family Independence Program)	1988-1993	Financial incentives and employment services	AFDC recipients
New York (Child Assistance Program)	1988-1994	Financial incentives and case management	AFDC recipients
New Chance	1989-1992	Educational, personal development, job training and employment services	Young high school dropout AFDC recipients

SOURCE: Greenberg and Shroder (1997).
NOTE: CWEP = Community Work Experience Program, AFDC = Aid to Families with Dependent Children. Randomized field trials (RFTs) are identified by their state of location, followed by their official name in parentheses, except in cases where the RFT took place in multiple states, which are identified only by their official name.
a. The experiment also tested the program on ex-addicts, ex-offenders, and young high school dropouts.

TABLE 2: Selected Randomized Field Trials of Reforms in Cash Welfare, 1960s to 1980s: Findings

	Findings
Negative income tax experiments	
New Jersey	Small negative effects on work effort of husbands, larger negative effects on work effort of wives
Rural	Significant reductions in family income and earnings, work effort of wives and dependents, little change in work effort of husbands
Seattle-Denver	Modest reductions in work effort of husbands, larger negative effects on work effort of wives and female heads
Gary	Significant reductions in work effort of husbands and female heads, no reductions for wives
National Supported Work Demonstration	Major positive effect on earnings
Maryland (Baltimore Options Program)	16% increase in earnings, no reductions in welfare expenditures
California (San Diego Job Search Demonstration)	Increases in employment rates that sometimes faded away; no effect on welfare recipiency rates
Washington (CWEP)	Increased employment rates
Arkansas (Work Demonstration)	Increases in employment and earnings, reductions in welfare recipiency and expenditures
Maine (Training Opportunities in Private Sector)	Positive effects on earnings but no effect on employment, welfare recipiency, or welfare expenditures
Virginia (Employment Services Program)	Small increases in employment and earnings
West Virginia (CWEP)	No effect on earnings, small reductions in welfare expenditures
AFDC Homemaker–Home-Health-Aide Demonstration	Generally positive effects on earnings and negative effects on welfare recipiency and expenditures

Florida (Trade Welfare for Work Program)	Mixed results on employment, earnings, and welfare expenditures
New Jersey (Grant Diversion Project)	Positive but fading effects on employment, positive effects on earnings, negative effects on welfare recipiency and expenditures
Illinois (Cook County Job Search and Work Experience)	No effects on employment and earnings but reductions in welfare expenditures
California (San Diego Saturation Work Initiative Model)	Positive effects on employment and earnings and negative effects on welfare recipiency and expenditures
Teenage Parent Demonstration	Modest positive effects on employment, mixed effects on earnings, negative effects on welfare recipiency
California (Greater Avenues for Independence)	Positive effects on employment and earnings and negative effects on welfare expenditures
Washington (Family Independence Program)	No effects on employment or earnings, positive effects on welfare recipiency and expenditures
New York (Child Assistance Program)	Positive effects on employment, no effects on welfare expenditures
New Chance	No effects on employment or earnings, positive effects on educational attainment, negative effects on emotional well-being

SOURCE: Greenberg and Shroder (1997).
NOTE: CWEP = Community Work Experience Program, AFDC = Aid to Families with Dependent Children. Randomized field trials (RFTs) are identified by their state of location, followed by their official name in parentheses, except in cases where the RFT took place in multiple states, which are identified only by their official name.
a. The experiment also tested the program on ex-addicts, ex-offenders, and young high school dropouts.

TABLE 3: Selected Randomized Field Trials of Reforms in Cash Welfare, 1990s: Design

	Random Assignment Period	*Description of Treatments*	*Population*
Florida (Project Independence)	1990-1993	Mandated job search, education, and training	AFDC recipients
Oklahoma (Oklahoma City NEWWS)	1991-1993	Education First, low enforcement of partial family sanctions	AFDC applicants
Michigan (Grand Rapids NEWWS Demonstration)	1991-1994	One treatment was Work First, second treatment was Education First; partial family sanctions in both	AFDC recipients and applicants
California (Riverside NEWWS)	1991-1993	One treatment was Work First, second treatment was Education first, partial family sanctions in both	AFDC recipients and applicants
California (Work Pays Demonstration Project)	1992	Enhanced earnings disregards, elimination of 100-hour rule, benefit level reduction	AFDC recipients
Georgia (Atlanta NEWWS)	1992-1994	One treatment was Work First, second treatment was Education first, partial family sanctions in both	AFDC recipients and applicants
Michigan (Detroit NEWWS)	1992-1994	Education First, low enforcement of partial family sanctions	AFDC recipients and applicants
Ohio (Columbia NEWWS)	1992-1994	Education First, high enforcement of partial family sanctions	AFDC recipients and applicants
New Jersey (Family Development Program)	1992-1994	Enhanced earnings disregards, strengthened sanctions, family cap	AFDC recipients and applicants
Canada (Self-Sufficiency Project)	1992-1995	Earnings supplement that can be taken off welfare	Single parent welfare recipients
Michigan (To Strengthen Michigan Families)	1992-1995	Enhanced earnings disregards, work requirements, partial family sanctions, elimination of 100-hour rule, diversion	AFDC recipients and applicants
Oregon (Portland NEWWS)	1993-1994	Work First or Education First assigned on basis of need; partial family sanctions	AFDC applicants and recipients

Iowa (Family Investment Program)	1993	Enhanced earnings disregards, work requirements, time limits, elimination of 100-hour rule, diversion	AFDC recipients and applicants
Vermont (Welfare Restructuring Project)	1994-1995	Enhanced earnings disregards, elimination of 100-hour rule, phased-in work requirements	AFDC recipients and applicants
Minnesota (Family Investment Program)	1994-1996	One treatment has enhanced earnings disregards only and a second treatment had enhanced earnings disregards combined with work requirements, sanctions, and elimination of the 100-hour rule	Long-term AFDC recipients and applicants
Wisconsin (New Hope Project)	1994-1995	Earnings supplements for those working at least 30 hours per week plus child care and health insurance	Low-income families
Florida (Family Transition Program)	1994-1995	Enhanced earnings disregards, work requirements, sanctions, time limits	AFDC recipients and applicants
Indiana (IMPACT Program)	1995	Enhanced earnings disregards, work requirements, sanctions, time limits, family cap, elimination of 100-hour rule, diversion	Job-ready AFDC recipients and applicants
Arizona (EMPOWER Program)	1995	Sanctions, time limits, family cap, elimination of 100-hour rule	AFDC recipients
Virginia (Independence Program)	1995	Enhanced earnings disregards, work requirements, sanctions, time limits, family cap, elimination of 100-hour rule, diversion	AFDC recipients
Delaware (A Better Chance Program)	1995	Enhanced earnings disregards, work requirements, sanctions, time limits, family cap, elimination of 100-hour rule, diversion	AFDC recipients and applicants
California (Los Angeles JOBS First-GAIN Program)	1996	Work First, partial family sanctions	AFDC recipients and applicants
Connecticut (Jobs First Program)	1996-1997	Enhanced earnings disregards, work requirements, sanctions, time limits, family cap, elimination of 100-hour rule	AFDC recipients and applicants

SOURCE: Grogger, Karoly, and Klerman (2002, Tables 3.4, 3.5).
NOTE: AFDC = Aid to Families with Dependent Children. Randomized field trials (RFTs) are identified by their location and by their official name in parentheses

TABLE 4: Selected Randomized Field Trials of Reforms in Cash Welfare, 1990s: Results

	Findings
Florida (Project Independence)	Modest positive impacts on earning but only in first year, and modest declines in welfare recipiency
Oklahoma (Oklahoma City NEWWS)	No impact on employment, earnings, household income, negative effect on welfare recipiency
Michigan (Grand Rapids NEWWS) Demonstration)	Positive impact on employment and earnings, negative impact on household income for some groups and treatments and no effect for others, negative impact on welfare recipiency
California (Riverside NEWWS)	Positive impact on employment, positive impact on earnings for one treatment and zero impact for the other, negative impacts on household income for some groups and no impact for others, negative impact on welfare recipiency
California (Work Pays Demonstration Project)	No impact on employment, earnings, household income, or welfare recipiency
Georgia (Atlanta NEWWS)	Positive impact on employment and earnings, no impact on household income, negative impact on welfare recipiency
Michigan (Detroit NEWWS)	Positive impact on employment and earnings, no impact on household income, negative effect on welfare recipiency
Ohio (Columbia NEWWS)	No impact on employment, positive impact on earnings, no impact on household income, negative effect on welfare recipiency
New Jersey (Family Development Program)	Negative impact on fertility
Canada (Self-Sufficiency Project)	Positive impact on employment, earnings, household income, and government transfers
Michigan (To Strengthen Michigan Families)	Positive impact on employment, earnings, and household income for some groups and no impact for others, negative impact on welfare recipiency

Oregon (Portland NEWWS)	Positive impact on employment and earnings, no impact on household income, negative impact on welfare recipiency
Iowa (Family Investment Program)	Positive impact on employment and earnings for some groups and not others and for some time periods and not others, generally positive impact on household income, little impact on welfare recipiency
Vermont (Welfare Restructuring Project)	Positive impact on employment or earnings for some treatments and not others, positive impact on household income in some data and not others and for some treatments but not others, no impact on welfare recipiency
Minnesota (Family Investment Program)	Positive impacts on employment and earnings for some groups, treatments, and time periods and none for others, generally positive effects on household income and welfare recipiency
Wisconsin (New Hope Project)	Positive impacts on employment, earnings, and household income for some groups and no impact for others, no impact on welfare recipiency
Florida (Family Transition Program)	Positive impacts on employment and earnings, some positive impacts on household income in short run, negative impact on welfare recipiency
Indiana (IMPACT Program)	Positive impact on employment and earnings, no impact on household income, negative impact on welfare recipiency in short run
Arizona (EMPOWER Program)	No impact on employment, earnings, household income, or welfare recipiency
Virginia (Independence Program)	Positive impact on employment and earnings, no impact on household income or welfare recipiency
Delaware (A Better Chance Program)	Positive impact on employment, no impact on earnings, household income, or welfare recipiency
California (Los Angeles JOBS First-GAIN Program)	Positive impact on employment and earnings, little impact on household income, negative impact on welfare recipiency
Connecticut (Jobs First Program)	Positive impact on employment and earnings, mostly no impact on household income, mixed positive and negative effects on welfare recipiency at different times

SOURCE: Grogger, Karoly, and Klerman (2002).

from the prior RFTs in several important respects. First, they were all tests of incremental reforms of the AFDC program, not structural reforms, because each tested some modification of the work requirements of the program within its then-existing structure. Second, the RFTs were all administered and conducted with the cooperation, and full partnership, of the AFDC agencies in each locality, unlike the NIT and Supported Work RFTs, which had set up separate operations to run the experiments outside of the existing AFDC system. Third, the RFTs were modest in scope and much less expensive than the other experiments and hence could be set up more quickly and more could be conducted for a given budget.

An important political development that affected the use of experimentation for evaluation began in the 1980s as well. Federal legislation in the early 1980s expanded states' ability to conduct tests of new reforms of their programs with the permission of the federal government, particularly statewide reforms. States were allowed to seek a waiver from federal law—that is, from the Social Security Act and its Amendments, which stipulated the requirements that all state AFDC programs must follow—for types of innovations and alterations of the program that had not been previously allowed. Many states were interested in testing the new reforms and this encouraged many to do so. At the same time, the federal government began to take an interest in the methods of evaluation used by the states to test the waiver reforms, with a strong belief that randomization should be the preferred method. Over the course of the 1980s, the federal government began encouraging states to use experiments as a method of evaluation, and in the late 1980s, the government persuaded many states to conduct experiments rather than use nonexperimental methods as part of the discussion granting waivers. Many states resisted randomization and often initially proposed nonexperimental evaluation designs, and it is unquestionable that many in this period would not have conducted RFTs without the federal requirement (Fishman & Weinberg, 1992). By the end of the 1980s and the beginning of the 1990s, the federal government was on the verge of making experimentation almost mandatory as a method of evaluation and a requirement before a waiver would be granted.

The importance of the federal ability to encourage RFT designs was even greater because the federal government no longer had the financial ability to pay the entire costs of new RFTs that it might deem valuable, as it did in the period of the NIT experiments, and hence, state cooperation was required. The NIT experiments cost more than $100 million in early 1970s dollars, including research as well as field expenditures. Federal budgets did not allow anything close to that kind of support by the 1980s, and hence, most RFTs cost in the range of $1 to $3 million (and sometimes less), with the state contributing considerable funds because the state often would pay for the reform innovation itself; the federal government only paid for the evaluation and analysis.[9]

The experiments of the 1980s did not have the design, attrition, or analysis weaknesses of the NIT experiments. The designs were in general extremely

simple, and the number of cells in the allocation program were kept small to keep the sample size in each large enough to preserve statistical power (generally only one experimental group, or at most two). Furthermore, the main outcome variables of interest were whether the programs increased work and earnings and decreased welfare participation rates, and data on these variables could be obtained from administrative records of the Unemployment Insurance and AFDC systems. These data were obtainable on all enrollees, including those who attrited from the experiment, and were not subject to underreporting. In addition, the analyses of the data were kept deliberately simple, consisting of simple experimental-control differences that were stratified by only a limited number of characteristics. Little or no attempts at behavioral modeling or estimation of responses to the program for endogenously defined groups (e.g., those who actually remained on welfare through the whole program) were made.

As a result of these strengths, the AFDC RFTs of the 1980s had great influence in state and federal policy circles. As Table 2 shows, the results were suggestive of gains in employment and earnings that could be had from modest investments in work-related programs for AFDC recipients. The results had a major impact on Congress and on the 1988 legislation mentioned above, which moved strongly toward work programs for AFDC recipients (Greenberg & Wiseman, 1992, pp. 61-62).

Generalization from the results of the 1980s experiments, or theoretical insights that might furnish the basis for generalization, were difficult and consequently scarce. The major attempt to deduce more generalizable lessons (Gueron & Pauly, 1991) attempted to group the various RFTs into broad-based (usually mandatory) programs with modest investments in simple job search or with other inexpensive programs, versus selective (usually voluntary) programs that provided more expensive education or training services on a smaller, more narrowly defined group. The study then attempted to assess which type was more cost effective. The experiments were not set up with this goal explicitly in mind, and consequently, the two groups of RFTs did not always have other characteristics held fixed, making the conclusions of the study rather problematic. In any case, however, subsequent policy has moved strongly toward the broad-based programs; therefore, the discussion of this difference has faded in prominence.

As noted above in the description of the history of welfare reform, welfare policy took a sharply different direction in the 1990s with the introduction of stronger work requirements, sanctions, and time limits. The RFTs followed this shift. Table 3 lists most of the major AFDC-related RFTs in the 1990s. All of these RFTs were begun in the pre-1996, pre-PRWORA period, a period in which states were eager to test alternative programs and interest in reform was accelerating.[10] Beginning in 1992, states who applied for waivers were granted them from the federal government only if they agreed, by and large, to randomization as a method of evaluation (U.S. DHHS, 1997), which applies to most

RFTs in the table that began in 1993 and after. As Table 3 indicates, these RFTs tested a variety of different reforms, ranging from Work First programs (which require that recipients become involved in a job search or employment activity immediately upon coming onto welfare) to Education First programs (which require the same immediate involvement in education or training programs) to work requirements, sanctions, time limits, and diversion (viz., imposing additional requirements on applicants that discourage them directly or indirectly from coming onto welfare). Enhanced earnings disregards and reduced tax rates—the hallmark of the NIT—also were prominent in many of these experiments, as were earnings supplements (wage or earnings subsidies, which have some similarity to enhanced earnings disregards but are somewhat different) (see Moffitt, 2003, for a comparison).

The scope of these RFTs lies somewhere in between the modest AFDC experiments of the 1980s and the larger scale experiments of the 1970s. Although they have tested reforms that are more far reaching than the incremental reforms tested in the 1980s, they fall far short of the radical reform tested by the NIT. In most other respects, however, they are similar to the RFTs of the 1980s. They were operated in cooperation with and administratively within the AFDC agencies of the locality, they were rather modest in cost (although in part because the reforms tended to reduce the caseload rather than increase it), they were simply designed and analyzed, and they utilized administrative data as the primary data source.[11] Consequently, they have the same potential for credibility and influence on policy as did the 1980s experiments.

Table 4 shows the main results of these RFTs to date (some are still releasing findings at this time). The majority of RFTs have shown positive effects on employment and earnings, suggesting that, indeed, reforms of the types begun in the early 1990s with work requirement, sanction, and other provisions could have an impact. The effects on household income are more ambiguous and mixed in sign for the simple reason that many families who left welfare because of these reforms experienced a loss in benefits, which largely cancelled their gains in earnings. Effects on welfare recipiency itself were negative in about half of the RFTs, suggesting a modest effect on the caseload, although many of the RFTs that most closely resembled post-1996 programs showed relatively little impact on welfare recipiency.

Although the potential for influence of the findings on policy is still present, the timing of the political process and the release of experimental results have not been as favorable as they were in the 1980s. The rush to reform in the early 1990s proceeded faster than results could be obtained, and consequently, the Congressional debate in 1996 and passage of the legislation took place at a time when many of the most important results shown in Table 4 had not been released. Nevertheless, discussions of the effects of the 1996 legislation have been widespread during the period since the law was passed and the findings from these RFTs have played a prominent role in that discussion, particularly on the relative impacts of Work First and Education First policies, for example, and

on the impact of financial incentives, as another. A limitation to the relevance of the RFT results, however, exists because they were all begun prior to 1996 and generally tested precursors to the programs have been ultimately implemented by the states subsequent to the 1996 law. Those precursors were not always, or even mostly, exactly like the post-1996 programs, which weakens the link between the RFT findings and current policies. This is a familiar issue with RFTs for experiments take time to conduct and to analyze, and policy interest often has proceeded some distance beyond the programs that were tested by the time the RFT findings are released. In any case, the ultimate influence of these RFTs on welfare legislation remains to be determined because the renewal or modification of the 1996 law has not yet been enacted by Congress at this writing.

There have been relatively few new RFTs begun since 1996. The primary reason is that the 1996 law devolved the program to the states and, hence, most federal regulatory authority disappeared as well because the states are no longer required to design programs according to any particular structure dictated by federal law. Consequently, there is no need for states to request waivers to test particular reforms and, hence, the prior use of waiver authority by the federal government as a means to require randomization is no longer available. Because many states are instinctively hostile to randomization for the usual reasons (e.g., perceived unethical nature of randomization; see Harvey, Camasso, & Jagannathan, 2000, for a discussion), they do not, as a general rule, use RFTs to evaluate their reforms if they evaluate them formally at all (there is no longer any requirement in the law to evaluate their programs and reforms). However, despite these barriers, the federal government has continued to work with states that are willing to conduct RFTs on subjects in which both the states and federal government have an interest. Thus, for example, an RFT evaluation of programs that can assist welfare recipients who have left the rolls to retain their jobs and advance in them is underway. Another RFT has been initiated that seeks to find programs that can assist those still on welfare who have the greatest employment barriers (the so-called hard to employ). Another RFT has begun in the area of child care, and others are under discussion. Nevertheless, although it remains to be seen how many states will use randomization as an evaluation method and how persuasive the federal government can be in this dimension, the number of RFTs in the future is likely to be less than in the period just before 1996.[12]

ASSESSMENT OF THE STRENGTHS AND WEAKNESSES OF RFTs IN THE AREA OF CASH WELFARE

The starting point of an assessment of the strengths and weaknesses of RFTs in the area of cash welfare must necessarily be one where their strengths are fully noted. Most of the RFTs in this area have been well conducted and

professionally analyzed. Randomization has, by and large, been conducted properly and maintained with integrity, with few problems of crossover, for example.[13] Internal validity has been, therefore, extremely strong. The advantages of simplicity of design, use of administrative data, simplicity of analysis methods, and policy relevance have led to a set of quite credible policy impact estimates. The significant influence of the experiments on the policy process properly reflects the strength of the methodology and the care and diligence with which RFTs have been implemented in so many different trials.

With these strengths taken as a given, however, the RFTs naturally also have limitations. Some, such as issues surrounding the ethics of randomization, are limitations that have been discussed many times before and apply to virtually all experimental evaluations. There have been numerous discussions within economics itself debating the importance of these general issues with randomized trials (e.g., Burtless, 1995; Heckman, 1992; Heckman & Smith, 1995). But there are several limitations with the cash welfare RFTs that are not mentioned in these general discussions or that take a very specific form in this area and that are partly unique to the historical and political circumstances of the cash welfare RFTs and the environment in which they have taken place. These are the limitations that will be discussed here, and there are five that will be covered. These are (a) contamination of control groups when estimating the effects of systemwide reform, (b) inability to estimate entry effects, (c) issues related to site effects, (d) limited and unplanned treatment variation, and (e) problems of black-box treatment designs.

Contamination of control groups when estimating the effects of systemwide reform. A familiar critique of RFTs is that they do not pick up the feedback or macroeffects that would occur if an experimental program were implemented nationwide (see Garfinkel, Manski, & Michalopoulos, 1992, for one discussion of this issue). A variety of such effects could occur. For example, those working through effects on markets, which economists term "general equilibrium" effects—a large increase in the supply of individuals to a particular labor market resulting from nationwide implementation, or a change in consumer demand resulting from increases or decreases in income—may change equilibrium wages or prices or unemployment rates, which will then feed back and alter the behavior of individuals in the population, generating an effect that is not captured by the small-scale RFT. "Going to scale," a term that is used to describe the process of going from a small pilot program to national implementation, is intended to capture these effects as well as effects that occur if the program undergoes alteration or changes its character when implemented on a large scale, or if the composition of recipients in the program is altered. Nevertheless, although these are all legitimate criticisms of RFTs, they pertain to the issue of externality validity—that is, generalizability of the findings from the RFT to other environments—and do not dispute their internal validity. The RFT is still

valid as an estimate of the first, initial impact of the program, before feedback effects occur and while the overall environment is unchanged.[14]

However, the cash welfare RFTs of the 1990s were vulnerable to a more serious manifestation of this problem because they took place in an environment in which such macroeffects were actually occurring and that almost certainly affected the outcomes of control group members. This is a more serious problem because it affects internal validity rather than external validity.

The reasons for feedback are numerous.[15] One key reason is that the nature of experimentation gradually shifted over the 1990s from small-scale to large-scale RFTs (U.S. DHHS, 1997). Although initially the RFTs were designed in the traditional way, with the experimental group composed of a randomly selected group of individuals small in size relative to the entire state recipient population, the designs gradually shifted over time to instead implement the new program on the entire state recipient population, except for a small randomly selected control group that was held back on the old program. With the entire state welfare population (excluding the small control group) on the new program, feedback and macroeffects are almost certain to occur and to affect the control group. Thus, this key shift in design fundamentally alters the inference that can be made from the experiments and threatens the validity of the results.

Although macro-, feedback effects are difficult to measure, they sometimes appeared in a very concrete form by an apparent confusion on the part of the control groups regarding which rules they actually faced. In several of the experiments, this "contamination" problem occurred as control group members, when interviewed and asked the rules facing them, mistakenly thought that they faced the rules of the new program (Camasso, Jagannathan, Harvey, & Killingsworth, 2003; Grogger, Karoly, & Klerman, 2002, p. 40; Harvey et al., 2000). Often, this occurred when publicity and media attention was devoted to the implementation of the new program statewide and control group members were exposed to that attention (Gordon, Jacobson, & Fraker, 1996).

In addition to this effect flowing from the changing design of RFTs, welfare reform gradually encompassed the entire nationwide caseload and changed the environment in such major ways that the control groups in the experiments were almost certainly affected. In the period prior to 1996, when states were given waivers to test and operate reform programs, more and more states took up this option until the majority of the nationwide caseload was no longer on the old AFDC program (Boehnen & Corbett, 1996; U.S. DHHS, 1997). By 1996, just prior to the passage of the Congressional legislation, more than 40 states had been granted waivers, for example. Many of the RFTs in Table 3 began in the first 3 or 4 years of the 1990s, and their early experimental-control comparisons may not have been affected by this change. However, RFTs in the later years, when outcomes continued to be measured, as well as those RFTs beginning in 1994 or later were almost certainly affected.

This situation worsened after 1996 when many of the RFTs in Table 3 were still being operated and outcomes for experimentals and controls were still

being compared. The PRWORA legislation had, it is now realized, a landmark, watershed effect on the low-income community. The transformation of the program from a pure cash program to a work program affected the perception of the program by those remaining on it as well as those in the low-income communities where large numbers of recipients or former recipients resided and the program became less of an option for those able to work. Entry rates into the TANF program declined dramatically and exit rates increased, resulting in a 50% drop in the caseload over the period 1994 to 1999. Attitudes toward work, childbearing, and other key behaviors in low-income communities have been drastically altered as common knowledge and expectations of the reforms have percolated through the families in those communities. Reform experts also believe that the stigma associated with being on welfare has increased as a social norm of work off welfare has taken hold and replaced the former norm of being on welfare and not working. Furthermore, the policy environment has been altered by welfare reform. Child care subsidies have increased by millions of dollars and new programs to assist low-income families with transportation to work have sprung up. Low-income communities and the helping agencies that proliferate within them have reoriented themselves to the new level of work among recipients and former recipients, and among those at risk of going onto welfare. These changes alone would affect the behavior of the control group members.

The lesson of this experience may be taken to be that small-scale RFTs, operated in a situation where the overall environment is unchanged, are best. It could be concluded, for example, that a more rational evaluation strategy would have been one in which the new reforms were tested on a small scale prior to the passage of a new law by Congress and prior to the statewide implementation that took place. However, it needs to be emphasized that policy makers involved in welfare reform very much desired to affect the overall environment by the welfare reforms of the 1990s. The explicit intent of the reforms was to change the culture of welfare and to change the message that low-income families were getting about welfare—that is, that it should now be all about work and about temporary rather than permanent assistance. The changes in perception within low-income communities, and the effects of those changes on individual educational, childbearing, and life decisions that have followed, were very much intended, or hoped-for, consequences of the reform. Any evaluation that did not capture those effects would, in the minds of program advocates, be missing a key effect of the reform. The proper lesson of the experience of the 1990s RFT of cash welfare reforms is instead that the RFT methodology is poorly suited to measuring the effects of structural, system-wide reforms that are intended to have as a large part of their effect the macro- and feedback effects from which it is impossible to insulate the control group.

Although it is hazardous to venture an estimate of the direction of the bias in RFTs created by this problem, the first presumption for this particular welfare reform should be that the RFT results should be biased downward because the control group also is affected by the reform. One piece of evidence supporting

this interpretation comes from a comparison of the RFT findings for the effect of reform on welfare recipiency and the welfare caseload, as compared to estimates of the impact of reform on those outcomes that come from nonexperimental, econometric estimates. The latter are based on studies that utilize time-series variation, comparisons of time series trends in outcomes for eligibles and ineligibles (i.e., so-called difference-in-difference designs), and related methods, all of which have their own pitfalls in internal validity. Nevertheless, although almost half of the RFTs in Table 4 showed either no effect or positive effects on welfare recipiency, the econometric studies almost uniformly show significant negative effects of reform (for reviews, see Blank, 2002; Grogger et al., 2002; Moffitt, 2003a). Moreover, even where the RFTs show negative effects on recipiency, their magnitudes are typically smaller than those evidenced in the econometric studies. This evidence is therefore consistent with the hypothesis that a decline in welfare recipiency among the control group in the RFTs could have been partially a result of the reform itself.

Inability to estimate entry effects. Entry effects occur when the implementation of a programmatic reform in an existing program alters the rate at which individuals apply for that program or gain admittance to it through the selection process of program operators. Reforms can have direct and indirect effects on entry. Direct effects can occur when the reform actually involves a change in the "front door" admission process by which applicants are handled. For example, the AFDC reforms of the 1990s involved, among other things, the introduction of diversion policies that were aimed at discouraging applicants from gaining entry to the program by offering them temporary payments to stay off welfare, requiring them to search for work prior to application, and other related policies. Indirect effects can occur when the policy reform affects the attractiveness of the program, in either a positive or negative direction, and consequently affects the rate at which eligibles apply. The indirect effects often grow over time as knowledge of the new reform percolates through the eligible population.[16] Measuring the effects of a reform on entry clearly requires estimates of the impact of that reform on the entry rate into the program in some way or another.

When studying the effect of the introduction of a new welfare program that has not been in existence, estimating their effects takes a slightly different form, namely, through their effects on the participation, or take-up, rate in the program. The total impact of a new program on a population logically has an overall effect that can be decomposed into two parts, one operating through the magnitude of the participation rate and one operating through the impact on the average outcomes of interest (e.g., earnings, employment, etc.) among those who choose to participate. The total impact on the average outcomes of the entire population is the product of these two variables, that is, the product of the fraction who take it up multiplied by the average outcomes of those who do. When studying a new program, generally it is just the overall participation

rate that is of initial interest, not necessarily the way in which entry and exit rates occur per se.

As discussed for the case of contamination and structural reform, one can distinguish between the separate issues of RFT designs that are small scale and those that are large scale. Traditional small-scale RFTs in the area of cash welfare, which randomize existing recipients, new recipients, or applicants into experimental and control groups, necessarily are incapable of estimating the impact of a reform on entry because the sampled population does not include those in the nonrecipient population who are at risk of entry and whose decisions may be altered (Moffitt, 1992b). RFTs to capture entry effects could be designed if the unit of observation were communities, or local areas, where a programmatic reform is offered in some areas and not others, because then the impact of the reform on the entry rate could be estimated by a comparison of that rate across experimental and control areas. Likewise, estimating the total effect of a new program would require that randomization take place over eligible populations, some of whom are offered the program and others not, and the participation rate is estimated by the experimental design. However, these types of area-unit designs have been judged to be infeasible in cash welfare and are rarely attempted (Hollister & Hill, 1995).

Nevertheless, the problem of inability to estimate entry effects and participation rates in small-scale RFT tests is a problem only of external validity because it implies that the generalization and extrapolation of the experimental results to a national program would provide an incomplete estimate of its total impact, much in the same way that macro-, feedback effects are missed.[17] Even if one could sample the nonrecipient population in the areas where a small-scale RFT has taken place, the fact that only a small group of recipients or applicants has been randomized into the experimental cell would almost surely have no effect on program entry because the eligibles would not perceive the experiment as having a sufficient impact on their own situation, because the probability of being selected for the experimental group, should they enter the program, would be negligible.

The small-scale RFTs of the 1980s and 1990s hence suffered from a problem of external validity arising from this source. Given the inability to design area-wide trials, nonexperimental estimates of the effects of the reform on program entry (e.g., either by cross-area comparisons of entry rates or those based on time series or difference-in-difference designs) are necessary to supplement the RFT findings and provide a more complete estimate of the total effect of the reforms on the outcomes of interest. This implies only that additional nonexperimental analyses would need to be added, but the experimental estimates are still valid for what they mean to accomplish.

Leaving out entry effects can nevertheless have a decisive effect on the interpretation of RFT results. This is best illustrated by the effect of increased earnings disregards or, equivalently, reduced welfare tax rates on work levels. Economic theory implies that a reduction in a welfare tax rate will have effects on

work effort that partly arise from changes in that effort among those who are initially welfare recipients and partly arise from changes in that effort among those not on welfare initially. The latter occur because tax rate reductions tend to draw new individuals onto welfare, an effect for which there is considerable research evidence. Although it is possible for the effect on work effort among initial recipients from the tax rate reduction to be positive, the effects from the new recipients who are drawn onto welfare are unambiguously negative. It is therefore possible for a recipient-only RFT design to show a positive effect of tax rate reductions even though the true, net effect is zero or negative.

A useful contrast to the recipient-only welfare RFTs of the 1980s and 1990s is provided by the NIT experiments, because those experiments differed from the later AFDC experiments by enrolling in the experimental and control groups a random sample of the entire low-income population in the area. Thus, individuals were enrolled who were not on welfare and who were, when randomized into the experimental group, offered the opportunity to enter the new welfare program if they wished but were not required to. Thus, entry effects—or, really, participation rate effects because this was a new program—were partially captured. Entry rate effects induced by changes in the welfare tax rate also were captured because the samples in each separate experimental cell included a random sample of the entire population, including nonparticipants, and thus, the effect of changes in the welfare tax rate on participation in the program could be estimated. This difference may be part of the explanation for why the estimated effects of welfare tax rate reductions in the NIT experiments on work levels and earnings showed no effects, whereas those of the 1980s and 1990s AFDC RFTs generally showed a positive effect (Moffitt, 2002, 2003a).[18]

However, it is questionable whether even the NIT experiments captured the same type of entry effects that would occur in a national implementation, because in those experiments individuals who did not participate were still enrolled in the experiment, had to submit monthly income reports and fulfill participation obligations in other respects, and most important, were explicitly and repeatedly informed of the programmatic options available to them. This would probably not replicate the information dissemination process in a national program, where there would be no such universal information availability or outreach. Participation and entry effects are likely, therefore, to be lower in a national implementation than in the NIT experiments. Thus, it is still likely that a design that randomizes across areas would, in principle, produce the best estimates of entry effects.

Issues related to site effects. The RFTs listed in Tables 1 and 3, including the NIT experiments, were conducted in a single area or a limited number of areas. The problem of external validity that this raises—that such areas may not be nationally representative and hence their results may not be a correct estimate of nationwide implementation—is a familiar one that has been discussed thoroughly in the literature. The problem arises if area-level characteristics interact

with the impact of the treatment on outcomes—as, for example, labor market characteristics are surely to do for programs aimed at affecting employment and earnings—and not so much if individual-level characteristics so interact. Variation in individual-level characteristics is available within individual sites where the RFTs have been conducted and, hence, interaction effects can be tested, in principle.[19]

The RFTs of the 1980s and 1990s are superior in this respect to those of the 1970s because a greater number have been conducted and in a much larger number of areas. Unfortunately, however, the key problem in learning the interactive effects of area characteristics on treatment impact is that the area variation embodied in the RFTs was not planned in any systematic way to provide variation from which something could be learned. Aside from variation in the treatments offered across areas, which confounds the cross-area interpretation of estimated impacts as resulting purely from area characteristics, the difference in area characteristics was not designed in such a way as to permit the estimation of the effect of variation in single area characteristics (e.g., the unemployment rate) holding fixed other characteristics (e.g., availability of other welfare and subsidy programs available, benefit levels in cash welfare, etc.). To estimate the effects of single area characteristics would have required that the areas be preselected in such a way, that is, to allow some area characteristics to be held fixed while others were varied.[20]

Although this type of planned variation is a common problem in all areas of social experimentation, the special political constraints discussed previously constitute a special barrier. Although the federal government had the regulatory authority to require random assignment as a method of evaluation in the period prior to 1996, it did not have the power to require states to conduct particular reforms. Initiation of a reform was a decision made by states themselves as they applied for waivers for particular changes they wanted to implement and to test. The set of area characteristics that resulted from the voluntary decisions of individual states, and the areas in which they proposed to test their reforms, was simply the set that fell out from which states submitted waivers. Although the federal government did have the power to suggest that particular areas within states be chosen for the evaluation, and design considerations often played an important role in their suggestions, there were clear limits in the variation that could result from this process of negotiation with the states. Thus, the type of planned variation that would have been needed was essentially impossible to achieve.

Two other issues in drawing inferences from the different areas in which RFTs of the 1980s and 1990s were conducted further demonstrate the potential seriousness of the problem. One concerns the characteristic of these RFTs that they generally enrolled only participants in cash welfare, an issue discussed already in the context of program entry. Welfare participation rates across states in the United States vary considerably, and not always because the individual characteristics of the eligible population (age, education, race, etc.) vary. They also vary because of the many different characteristics of the AFDC programs in

the different states, including benefit levels, asset tests, prereform work programs, and a variety of other program rules and characteristics. In addition, even holding these rules fixed, differences in the stigma of being a welfare recipient, and variations in social norms, both of which are difficult to measure and hence are essentially unobserved, affect take-up. This implies that the composition of the recipient populations varies across states in the United States in unobserved ways that are likely to interact with the impact of reform programs on employment, earnings, and related outcomes. This complicates the inferences that can be drawn from cross-area comparisons because estimates of program impact may differ only because the composition of the recipient population differs.

The second issue concerns the variation in estimated effects within individual RFTs in different local welfare offices or different sites. Many of the RFTs conducted their treatments in a number of such offices and sites within the general experimental area, and in most cases, there was significant variation in the estimated treatment impact (Greenberg, Meyer, Michalopoulos, & Wiseman, 2001). This occurred even though the treatment was intended to be basically similar across offices and sites. Generally, these office and site effects cannot be explained adequately by any measurable variable. It is possible that the treatment was in fact implemented differently in different areas, but it is just as likely that the composition of the populations was different or that there were area-specific characteristics that were strongly interacting with the treatment. These results suggest that the generalizability problem, and the problem of inferring the effects of area characteristics on treatment impacts, is a serious one that constitutes a significant limitation of the designs.

Even with all of these problems, there were a sufficient number of different areas and sites in the RFTs of the 1980s and 1990s that meta-analyses can be, and have been, conducted (Bloom, Hill, & Riccio, 2001; Greenberg, Ashworth, Cebulla, & Walker, 2003; Greenberg et al., 2001). The dependent variable in such a meta-analysis is generally the estimated treatment impact for a particular program in a particular area, and the independent variables are the characteristics of the program and the characteristics of the area.[21] The major problem with the meta-analyses is that the effective sample size—that is, the number of areas in which experiments have taken place—is still too small to disentangle the separate effects of site and treatment effects or too small to represent more than one or two site characteristics. The Greenberg et al. analysis, for example, showed that when a minimal set of all area characteristics and treatment characteristics was included in the model, the coefficients on virtually all important programmatic characteristic variables became statistically insignificant. In the Bloom et al. analysis, only one area characteristic—the unemployment rate—was entered, which does not adequately capture the differences across areas that would need to be captured to generalize to a national estimate.

Limited and unplanned treatment variation. A rather related issue that, again, pertains to the ability to use RFTs to learn lessons for the future is the extent to

which the RFTs of the 1980s and 1990s reflected limited and unplanned treatment variation. The analogous problem to area variation discussed previously was present in this case as well, because the programs tested in different areas did not vary particular treatment features while holding others fixed, thus preventing learning the incremental effects of particular treatment components. Once again, the political constraints on the constellation of RFTs that were conducted is one of the primary reasons for this lack of variation.

A special feature of the 1980s and 1990s RFTs, however, was the extent to which the treatments offered were complex bundles, or packages, of multiple reform components. For example, a particular reform might involve the imposition of time limits, work requirements, a particular type of sanction policy, a family cap, a certain level of earnings disregards, and perhaps minimum hours requirements for receipt of benefits. This type of bundled reform was gradually adopted by states throughout the 1990s. In the early years of that decade, states tended to be interested in testing one or two components at a time, but they later moved toward testing multiple components (Boehnen & Corbett, 1996). This was an intentional policy shift because policy makers were most interested in changing the entire nature of the welfare system and this meant changing many components at the same time. With this type of bundling in the RFTs, the challenge to learn the effects of each individual component was particularly great, because it would have required conducting a relatively large set of RFTs that held multiple components of the bundle fixed while varying others.[22] This is not an inherent barrier of the RFT methodology and, indeed, the RFT methodology is in many respects ideally suited to estimating the incremental effects of alternative program components. However, the political and bureaucratic constraints on doing so prevented such planned designs from taking place.

The experience of the RFTs of the 1980s and 1990s in this respect poses a political difficulty for experimental design if the estimation of the incremental effects of individual components are of interest. The policy makers over this period were initially not interested in testing the effects of individual treatment components added on top of the then-existing AFDC program. This is because the policy makers believed that the effects of the individual components interact and that the sum total effect of the bundle as a whole would be greater than the effects of any individual component-introduced piecemeal. Once again, it was the effect of transforming the program in a major way that was the object of interest. However, once the effects of such bundled reforms have been estimated, it is likely that policy makers will be interested in learning the effects of adding or subtracting, or altering the nature of, individual components in the bundle in the future, starting from the base of having already implemented one major program bundle (the incremental effects of such reforms are no doubt different than those that would have been estimated starting from the AFDC program as a base). This suggests that a sequential RFT strategy may be optimal, starting with a bundled reform and then later proceeding with incremental reforms on top of it. However, to carry this out requires that the policy makers

involved have a sustained commitment to sequential RFT formulation that, at least in the area of social welfare, has been absent. The second stage has been difficult to generate political support for because programs have devolved to the states and randomization is no longer a requirement for receiving federal dollars. In retrospect, therefore, it would have been better to have built in some component variation from the beginning.

The rather large number of areas in which bundled treatments were tested makes, once again, a meta-analysis possible, which could, in principle, indirectly estimate the increment effects of reform components by comparisons across RFTs, holding constant other differences. However, the degrees of freedom necessary to estimate those incremental effects, when they are known to interact and therefore depend on the initial bundle in place, limit the extent to which this can be achieved. Greenberg et al. (2001), for example, showed that this is infeasible given the existing number of RFTs and the amount of variation that needs representation in the model.[23]

A few of the RFTs in Table 3 did, however, introduce variation in treatments within RFTs and thus were able to compare alternative policies holding area characteristics fixed. One example came from the tests of Work First versus Education First treatments, which were conducted in several areas. Another was the variation in treatments that offered new financial incentives (e.g., reduced earnings disregards) alone and those that offered those financial incentives in addition to some type of work requirement. The results of these comparisons have in fact turned out to be very valuable in policy discussions for this reason and the body of knowledge on the relative effects of these program components is one of the strongest set of findings to come out of the RFT literature. But these treatment comparisons are the exception rather than the rule, and most other components tested in the RFTs have not experienced such direct variation.[24]

Problems of black-box treatment designs. The final issue is that of black-box treatment designs. Black-box treatments are those constituted of multiple complex treatment components that are either difficult to describe or that allow considerable discretion when implemented in the field. A welfare-to-work program, for example, which consists of some type of initial assessment of job skills, assignment to a type of work or training program for which the caseworker is given general guidelines but allowed discretion, followed by a sequence of work programs and sanctions, the latter of which is also partly at the discretion of the caseworker, is a case in point. The treatment is composed of multiple stages, discretion is allowed, and the exact nature of the treatment given each individual is not spelled out in the experimental protocol. The term "black box" refers to the fact that the actual treatment in the RFT, which transforms the outcome variables of individuals from their preexperimental values to their postexperimental values, takes place inside of a "box" that is shielded from view inasmuch as it is not easily understood and characterized.[25]

Black-box treatments also can be understood by comparison with their polar opposite, which is a simple treatment that is fully quantified and therefore characterizable in measurable terms. A simple alteration of a benefit level, for example, would fall into this category. The treatment is transparent and easily understood.

Some analysts also regard black-box treatments as those where the mechanism by which the treatment has an effect is not understood or where there is no theory to guide the experiment (where a theory in this case is a hypothesis about the mechanism by which the treatment affects outcomes). However, this is a less fundamental distinction, because RFTs with quantifiable and easily characterized treatments, strictly speaking, can be informative even if the mechanism by which the treatment affects outcomes is not fully known. Indeed, RFTs in general, similar to much nonexperimental work, are not informative on mechanisms.

The problems that black-box experiments raise are, first, that they are difficult to replicate and, by extension, difficult to generalize to a national program; second, it is difficult to compare different black-box experiments to each other or to extrapolate from them to programs that may differ from them in small or large ways. RFTs with treatments that are difficult to characterize and that allow room for local variation and discretion may not be replicatable, and this has indeed occurred in the cash welfare RFT experience where some RFTs that showed outstanding positive effects could not be replicated in other areas.[26] The same features of the treatment in an experiment could render hazardous the generalization to a nationwide program where implementation could be quite different. The problem of comparisons across different black-box RFTs is also readily understood. It is generally difficult to know what to make of differences in impact estimates between two RFTs that each have complex and difficult-to-characterize treatments. Extrapolation is difficult as well because it is virtually impossible to know whether the alteration of any one of the dozens of small, individual components of a black-box treatment would have a large or small effect on outcomes. This relates to national implementation as well because such implementation would almost surely result in changes in some individual components of the treatment.

The black-box problem in cash welfare RFTs compounds the problem of bundling described previously. If a treatment has 10 major individual components that are bundled together and each of the 10 itself is a black-box treatment, then the difficulties in replication, generalization, and extrapolation are that much larger. The goal of conducting planned variation in major individual reform components to learn the incremental effects of each is difficult to conduct when those components are black box in nature.

The solution to the black-box problem is to establish a characterization, or typology, of treatments that consists of a relatively small number of building blocks built up from the major elements of the approach. Although it would be preferable if the elements of the typology could be cardinalized, and therefore

quantified, an ordinal typology—for example, one that ranks treatments along a single dimension or small number of dimensions as "weak" or "strong"—also would be an advance over the typical black-box design. This would permit different RFTs to be compared because the building blocks could be compared, as could their ordinal or cardinal rankings. Treatments would need to be described in detail to eliminate discretion in their implementation, which would make it more likely to achieve the goals of generalization and replication. With such a typology in place, the type of planned variation described earlier in which the building blocks and their ordinal or cardinal rankings are systemally varied, holding others fixed, could take place.

The NIT experiments again can be usefully contrasted with the RFTs of the 1980s and 1990s in this respect. The NIT experiments were at the polar opposite of black-box experiments because they were explicitly intended to estimate not the effects of any particular specific reform program but rather the effects of variation in guarantees and tax rates on work levels. The explicit intention of the designers of the NIT experiments was to have a sufficiently wide variation in guarantees and tax rates across the different treatments so as to be able to reliably estimate the response surface, that is, the slope of the line (determining outcomes) with respect to guarantees and tax rates. The designers even went so far as to pick the sample sizes across different guarantee-tax-rate treatments to minimize the variance of an estimated regression coefficient for such a slope. Although it has to be recognized that any such estimates have to involve interpolation and extrapolation beyond what the point estimates of the experiment can directly provide from its finite number of experimental cells, it is nevertheless the case that the conceptual framework brought to bear was a powerful one that at least allowed the possibility that something might be learned from the experiment beyond the specific treatments tested. Forecasting the effects of new programs was built into the design from the beginning. These considerations were absent from the RFTs of the 1980s and 1990s.

CONCLUSIONS

The RFTs in the area of cash welfare in the past 30 years have produced much valuable and credible information on the effects of various reforms, information that has often properly played a major role in the policy process. However, the five limitations of those RFTs described in this article, some of which are inherent to the RFT design and others of which are a result of economic or political constraints, circumscribe what has been learned from past RFTs and what is likely to be learned from future RFTs in this area.

Each of the five limitations has a corresponding lesson. The lesson of the first limitation is that RFTs are best used when they attempt to estimate the effect of incremental reform within a given, overall programmatic structure and are poorly designed to estimate the effect of systemwide, structural reform that

alters the entire environment. Estimates of the latter should be reserved for nonexperimental analyses. All indications are that the U.S. social welfare system is in for a period of relative overall stability for several years because the general structure of the 1990s reforms is very popular among the voters and members of Congress. Consequently, the odds that a major change will be adopted in the near future are low. This would seem to imply that there should be ample productive opportunities for small RFTs that test incremental reforms and that search for specific, detailed policies that reveal "what works and for whom."

Second, RFTs should be supplemented by nonexperimental analyses of entry effects where it appears possible that those effects are significant. Although RFTs that offer a random set of individuals in a location a program reform that they are not obliged to take are possible, this often does not replicate the information mechanisms about the reform that would take place in a national program in which nonparticipating eligibles are not directly told of the reform. Moreover, the scientific and political constraints to randomizing across areas, which is the best way to estimate entry effects, are sufficiently severe that it is unlikely that they will ever be used in a systematic way to estimate those effects. Therefore, RFTs should be reserved for estimating the exit effects and effects on initial participant populations.

The other three limitations discussed in this article all pertain to external validity and are concerned with ways to learn more about policy alternatives than recent RFTs have been able to do. Planned, systematic variation in both area characteristics and program characteristics, based on the idea of varying one characteristic or group of characteristics across different RFTs—or within a single RFT—while holding other characteristics fixed, constitutes the basic design that is necessary. Political constraints are the major barrier to carrying out such a set of planned RFTs. In the absence of such variation, nonexperimental analyses are needed to extrapolate from RFTs, analyses that necessarily have to make identifying assumptions that are not based on randomization. Finally, a systematic typology that quantifies the intensity and character of the various individual components of welfare interventions needs to be developed and applied in the design of social welfare RFTs.

NOTES

1. See Moffitt (2003) for a detailed review of this history.

2. For a brief history of the negative income tax idea and its fate in U.S. welfare policy, see Moffitt (2003b). See Gueron and Pauly (1991) for a detailed discussion of the shift that occurred in the 1980s toward work-requirement programs.

3. Sanctions were introduced in 1971 but did not grow to significant levels nationwide until the 1990s.

4. See Greenberg and Shroder (1997) for a complete listing of those prior to the 1990s and Grogger, Karoly, and Klerman (2002) for a complete listing of those in the 1990s. Tables 1 and 2 omit the

Work Incentive Program (WIN) laboratory experiments and a number of other Aid to Families with Dependent Children (AFDC) experiments that were generally smaller in scope and depth than those shown in Table 1. Also, see Greenberg and Shroder (1997) and Hausman and Wise (1985a) for reviews and listings of social welfare randomized field trials (RFTs) in areas other than cash welfare.

5. See Moffitt (1992a, 1992b, 2002, 2003a, 2003b) for discussions of why simple economic theory predicts the possibility of no effect. As it turned out, the negative income tax (NIT) experiments yielded estimates of the effects of changes in guarantees and tax rates that were within the range of, and consistent with, nonexperimental estimates of the effects of those variables (Moffitt & Kehrer, 1981) and therefore were accurately forecastable from the prior nonexperimental evidence.

6. This finding was directly inferred from the experimental-control difference for married men and women because the control group essentially faced no welfare program; for single mothers, the experimental guarantees were set above those for the AFDC program faced by the control group, which also led to an experimental-control difference that measured the effect of increasing the generosity of benefits. The effect of tax rates per se, which was of less Congressional interest, was measured by differences in work effort between cells of the experimental design specifying the same guarantee but different tax rates. See Moffitt (2002a) for a summary of the experimental evidence on this score.

7. A structural model in this context is one that uses economic theory to specify the response equation and enters the experimental-control dummy on the right side not as a simple linear, additive effect but rather interacted with other variables and often in a nonlinear form as suggested by theory.

8. Two other problems (i.e., the restriction of the sample to families with incomes below a specified level, which causes what is known as a truncation problem because it picks up families whose incomes are only temporarily low, and the limited duration of the experiments, only a few years in length, which could have induced behavior different than what would occur in a permanent program) are also not inherent, although subsequent experiments have not addressed them in any better fashion. Most current experiments use AFDC status at a point in time to define the sample on which randomization is conducted, which will result in a disproportionate number of long-term recipients and will miss others who happen not to be on the rolls at that point in time; in addition, most experiments are still of limited, or at least uncertain, duration.

9. However, there were federal matching funds available for the administrative costs of state experiments that came out of general federal entitlement funds for the program, not out of the federal research budget, which relaxed the budget constraint somewhat.

10. The National Evaluation of Welfare to Work Strategies (NEWWS) evaluations in Table 3, although begun in the 1990s, were planned in the 1980s and were intended to evaluate the work reforms embodied in the 1988 federal legislation referred to previously. They are included in Table 3 rather than in Table 2 because many of their features are similar to those of the other 1990s RFTs.

11. Occasionally an RFT was operated outside the existing AFDC agency (e.g., the New Hope Demonstration).

12. The continued limited nature of federal budgets, referred to earlier, also implies that the federal government does not have the financial ability to simply pay for RFTs that it might be interested in but which no state is. Thus, any RFT to be conducted has to be one in which the state has an interest. This will be discussed further below.

13. See Harvey, Camasso, and Jagannathan (2000) for a discussion, who note that there is no evidence that social workers violated the integrity of random assignment even though they were averse to denying treatment. Cross-over problems are not always fully documented, however. For example, Gordon, Jacobsen, and Fraker (1996) note that some of the waiver RFTs of the early 1990s did not have sufficient tracking systems to know when cross-over occurred. But the authors concluded from their select examination of a few RFTs that cross-over was nevertheless not large in magnitude.

14. Moreover, the criticism applies equally to many nonexperimental estimators, which likewise capture only the partial-equilibrium and not the general-equilibrium impacts of programs.

15. See Moffitt and Ver Ploeg (2002) for a prior discussion.

16. See Moffitt (1996) for a more detailed discussion of the different types of entry effects in the context of the effect of the introduction of work-related program into cash welfare. See Grogger, Haider, and Klerman (2003) for evidence that the 1990s welfare reforms had a large effect on entry.

17. A slightly more subtle problem of generalization occurs if a change in the entry rate upon nationwide implementation would bring onto the program (or force off) individuals who are different than those enrolled in the RFT in terms of the outcome variables of interest. This implies that nationwide implementation could result in a caseload for which the average impact of the new programmatic reform could differ from that in the experiment (Heckman, 1992; Moffitt, 1992b).

18. Another type of entry effect that is estimable with experimental data is the effect of offering a subprogram within welfare to existing welfare recipients. Welfare recipients can be randomized into those who are offered the subprogram and those who are not, and takeup of the subprogram can be estimated by experimental-control differences. See Card, Robins, and Lin (1997) for an analysis of an RFT that contained a design for this type of effect.

19. If either area-level or individual-level characteristics affect outcomes similarly and additively for experimentals and controls, this does not cause a problem and experimental-control differences are still comparable across RFTs.

20. In the literature on experimental design, such designs are called factorial designs because they generally do not propose a set of experimental groups with every possible combination of characteristics but rather only a selected set with some assumed additivity in the effects of each (i.e., as in a factor model). Assumptions about additivity are a necessary part of such designs.

21. Occasionally, a study returns to the individual data and reestimates the treatment effects as part of the meta-analysis (e.g., Bloom, Hill, & Riccio, 2001). See also Hotz, Imbens, and Mortimer (1999; Hotz, Imbens, & Klerman, 2000) for other recent attempts to compare results across experimental sites.

22. As noted previously, to avoid testing all possible combinations, factorial designs would have been the preferred strategy.

23. As Greenberg, Meyer, Michalopoulos, and Wiseman (2001) note, there have been a number of more informal efforts to draw lessons from multiple experimental results (e.g., Bloom & Michalopoulos, 2001). These efforts typically make judgments about which components of the different treatments are most important and what one or two area characteristics are important and then attempt to cross-classify the RFTs, post hoc, into two or three treatment types and, possibly, one or two area types. Although such an analysis is often informative, the Greenberg et al. analysis shows that an expansion of the number of treatment and area characteristics renders such studies lacking in power and statistical significance. It also should be noted that some true meta-analyses (e.g., Bloom et al., 2001; Greenberg, Ashworth, Cebulla, & Walker, 2003) do not attempt to characterize the entire bundle of treatment components but rather select only a subset to include in the analysis.

24. Once again, the political constraint that essentially requires any treatment variation to be one in which the state itself has an interest limits the types of variation that could be tested. For example, thus far, no state has been particularly interested in offering programs without a time limit—but with the other components such as work requirements and sanctions held fixed—because states are already convinced of the desirability of time limits. Thus, there is no RFT evidence on the effects of time limits per se, even though this would have been an eminently appropriate policy whose effects could be estimated with the RFT methodology.

25. The black-box problem is not the same as the bundling problem referred to above. For example, an experiment that simultaneously altered two or more parameters of a welfare benefit formula would encounter the bundling problem of disentangling their separate effects, but not the black-box problem because the components are transparent, easily quantified, and replicable in other areas and times. It is the difficulty in characterizing and standardizing the treatment that is the essential feature of the black-box problem, which could take place even if there is only one treatment component.

26. Programs in Riverside, California, and Portland, Oregon, for example, showed much greater effects than other programs with apparently similar characteristics and have not been fully explained. The Riverside program results were partly replicated in a later RFT in Los Angeles, but Greenberg et al. (2003) are not able to account for much of the difference between the Riverside and Portland effects and those of other experiments.

REFERENCES

Ashenfelter, O., & Plant, M. (1990, January). Nonparametric estimates of the labor-supply effects of negative income tax programs. *Journal of Labor Economics*, *8*, S396-S415.

Blank, R. (2002, December). Evaluating welfare reform in the United States. *Journal of Economic Literature*, *4*, 1105-1166.

Bloom, D., & Michalopoulos, C. (2001). *How welfare and work policies affect employment and income: A synthesis of research*. New York: Manpower Demonstration Research Corporation.

Bloom, H., Hill, C., & Riccio, J. (2001). *Modeling the performance of welfare-to-work programs: The effects of program management and services, economic environment, and client characteristics*. New York: Manpower Demonstration Research Corporation.

Boehnen, E., & Corbett, T. (1996). Welfare waivers: Some salient trends. *Focus*, *18*(1), 34-37.

Burtless, G. (1987). The work response to a guaranteed income: A survey of the experimental evidence. In A. Munnell (Ed.), *Lessons from the income maintenance experiments*. Boston: Federal Reserve Bank and Brookings.

Burtless, G. (1995, spring). The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives*, *9*, 63-84.

Camasso, M., Jagannathan, R., Harvey, C., & Killingsworth, M. (2003). The use of client surveys to gauge the threat of contamination in welfare reform experiments. *Journal of Policy Analysis and Management*, *22*, 207-233.

Card, D., Robins, P., & Lin, W. (1997). *Would financial incentives for leaving welfare lead some people to stay on welfare longer? An experimental evaluation of "entry effects" in the self-sufficiency project*. New York: Manpower Demonstration Research Corporation.

Fishman, M., & Weinberg, D. (1992). The role of evaluation in state welfare reform waiver demonstrations. In C. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs*. Cambridge, MA: Harvard University Press.

Friedman, M. (1962). *Capitalism and freedom*. Chicago: University of Chicago Press.

Garfinkel, I., Manski, C., & Michalopoulos, C. (1992). Micro experiments and macro effects. In C. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs*. Cambridge, MA: Harvard University Press.

Gordon, A., Jacobson, J., & Fraker, T. (1996). *Approaches to evaluating welfare reform: Lessons from five state demonstrations*. Washington, DC: Mathematica Policy Research.

Greenberg, D., Ashworth, K., Cebulla, A., & Walker, R. (2003). *When welfare-to-work programs seem to work well: Explaining why Riverside and Portland shine so brightly*. Mimeographed, University of Maryland at Baltimore County.

Greenberg, D., Meyer, R., Michalopoulos, C., & Wiseman, M. (2001). *Explaining variation in the effects of welfare-to-work programs*. Madison, WI: IRP DP 1225-01.

Greenberg, D., & Shroder, M. (1997). *Digest of social experiments*. Washington, DC: Urban Institute.

Greenberg, D., & Wiseman, M. (1992). What did the OBRA demonstrations do? In C. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs*. Cambridge, MA: Harvard University Press.

Grogger, J., Haider, S., & Klerman, J. (2003, May). Why did the welfare rolls fall during the 1990s? The importance of entry. *American Economic Review*, *93*, 288-292.

Grogger, J., Karoly, L., & Klerman, J. (2002). *Consequences of welfare reform: A research synthesis*. Washington, DC: DHHS.

Gueron, J., & Pauly, M. (1991). *From welfare to work*. New York: Russell Sage.

Harvey, C., Camasso, M., & Jagannathan, R. (2000, fall). Welfare reform evaluation under section 1115. *Journal of Economic Perspectives*, *14*, 165-188.

Hausman, J., & Wise, D. (1985a). *Social experimentation*. Chicago: University of Chicago Press.

Hausman, J., & Wise, D. (1985b). Technical problems in social experimentation: Cost vs. ease of analysis. In J. Hausman & D. Wise (Eds.), *Social experimentation*. Chicago: University of Chicago Press.

Heckman, J. (1992). Randomization and social policy evaluation. In C. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs*. Cambridge, MA: Harvard University Press.

Heckman, J., & Smith, J. (1995, spring). Assessing the case for social experiments. *Journal of Economic Perspectives*, *9*, 85-110.

Hollister, R., & Hill, J. (1995). *Problems in the evaluation of community-wide initiatives*. New York: Russell Sage.

Hotz, V., Imbens, G., & Klerman, J. (2000). *The long-term gains from gain: A re-analysis of the impacts of the California GAIN program* (Working Paper 8007). Cambridge, MA: National Bureau of Economic Research.

Hotz, V., Imbens, G., & Mortimer, J. (1999). *Predicting the efficacy of future training programs using past experiences* (Technical Working Paper 238). Cambridge, MA: National Bureau of Economic Research.

Lampman, R. (1965, May). Approaches to the reduction of poverty. *American Economic Review*, *55*, 521-529.

Lampman, R. (1968). *Expanding the American system of transfers to do more for the poor: U.S. Congress, Joint Economic Committee.* Washington, DC: Government Printing Office.

Moffitt, R. (1992a). Evaluation methods for program entry effects. In C. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs*. Cambridge, MA: Harvard University Press.

Moffitt, R. (1992b, March). Incentive effects of the U.S. welfare system: A review. *Journal of Economic Literature*, *30*, 1-61.

Moffitt, R. (1996, winter). The effect of employment and training programs on entry and exit from the welfare caseload. *Journal of Policy Analysis and Management*, *15*, 32-50.

Moffitt, R. (2002). Welfare programs and labor supply. In A. Auerbach & M. Feldstein (Eds.), *Handbook of public economics* (Vol. 4). Chicago: University of Chicago Press.

Moffitt, R. (2003a). The temporary assistance for needy families program. In R. Moffitt (Ed.), *Means-tested transfer programs in the U.S.* Chicago: University of Chicago Press.

Moffitt, R. (2003b, Summer). The negative income tax and the evolution of U.S. welfare policy. *Journal of Economic Perspectives*, *17*, 119-140.

Moffitt, R., & Kehrer, K. (1981). The effect of tax and transfer programs on labor supply: The evidence from the income maintenance experiments. In R. Ehrenberg (Ed.), *Research in labor economics* (Vol. 4). Greenwich, CT: JAI.

Moffitt, R., & Ver Ploeg, M. (Eds.). (2002). *Evaluating welfare reform in an era of transition*. Washington, DC: National Academy Press.

Moynihan, D. (1973). *The politics of a guaranteed annual income.* New York: Vintage.

Pechman, J., & Timpane, P. M. (Eds.). (1975). *Work incentives and income guarantees*. Washington, DC: Brookings Institution.

SRI International. (1983). *Final report of the Seattle/Denver income maintenance experiment. Volume I: Design and results*. Menlo Park, CA: Author.

Tobin, J. (1966, fall). On the economic status of the Negro. *Daedalus*, pp. 89-895.

Tobin, J., Pechman, J., & Mieszkowski, P. (1967, November). Is a negative income tax practical? *Yale Law Journal*, *77*, 1-27.

U.S. DHHS. (1997). *Setting the baseline: A report on state welfare waivers*. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation.

ROBERT A. MOFFITT *is professor of economics at Johns Hopkins University. His research has concerned welfare reform, program evaluation methodology, and statistical methods. He is the coeditor of the National Research Council volume* Evaluating Welfare Reform in an Era of Transition.