Using Experiments to Evaluate Performance Standards

What Do Welfare-to-Work Demonstrations Reveal to Welfare Reformers?

John V. Pepper

ABSTRACT

This paper examines how experimental demonstrations can be used to inform planners about the efficacy of social programs in light of a performance standard. The problem is illustrated by considering the situation faced by state governments attempting to design programs to meet the new federal welfare-to-work standards. Data from experimental evaluations alone allow only limited inferences about the labor market outcomes of welfare recipients. Combined with prior information on the selection process, however, these data are informative, suggesting either that the long-run federal requirements cannot be met or that these standards will only be met under special circumstances.

I. Introduction

In August 1996, the open-ended federal entitlement program of Aid to Families with Dependent Children was replaced with the Temporary Assistance for Needy Families (TANF) block grant for states. The Personal Responsibility and

John V. Pepper (jvpepper@virginia.edu) is an assistant professor of economics at the University of Virginia. *This research has been supported in part by the 1998–99 Joint Center for Poverty Research/ASPE Small Grants Program. The author has benefitted from the comments of Stephen Bell, Dan Black, Leora Friedberg, Carolyn Heinrich, V. Joseph Hotz, Matt Lyon, Charles Manski, and three anonymous referees. The author also has benefitted from the opportunity to present this work at the 1999 JCPR Small Grants Conference, the 2000 ASSA Conference, the 2000 Murray S. Johnson Memorial Conference at the University of Texas, and in seminars at Colorado University, Cornell University, the General Accounting Office, Mathematica, Northwestern University, Syracuse University, and the University of Virginia. He thanks Jessica Howell for research assistance. The data used in this paper are derived from files made available to researchers by the MDRC. The author remains solely responsible for how the data have been used and interpreted.*
[Submitted April 2000; accepted May 2002]
ISSN 022-166X

Work Opportunity Reconciliation Act of 1996 (PRWORA) gives states broad discretion in defining eligibility and benefits for welfare programs. TANF funds, however, are not guaranteed. For state governments to be assured of continued full funding, at least 25 percent of single parent recipients must be participating in work activities in 1997, with this proportion increasing by 5 percent each year until it reaches 50 percent in 2002. Thus, under the new law state governments must design and implement welfare programs that both provide assistance and encourage work, two objectives that have so far appeared incompatible.

Planners often face the difficult task of designing and implementing social programs that meet outcome objectives or performance standards. In this setting, randomized experiments on innovative programs are sometimes thought to provide relevant information. Over the past four decades, hundreds of randomized experiments have been used to evaluate the effectiveness of a variety of social programs including education, energy pricing, health insurance, housing subsidies, illicit drug policies, labor market training, policing, prisoner rehabilitation, and welfare reform. For many innovative programs, these experiments may be the only source of observed data.

What do these data reveal? Consider, for instance, the problem of using the welfare-to-work demonstrations to inform state policy makers. Typically, a sample of welfare recipients is randomly placed in either a treatment group which receives job training or a control group which receives the standard benefits. At some point in the future, the labor market outcomes of each individual are observed. Thus, in principle, the sampling process identifies the fraction of welfare recipients who would work if they were all to receive training or if instead they were all to receive the standard benefits. By comparing these probabilities, researchers measure the effectiveness of the training program.

Although experiments yield information on the outcomes of mandated treatments, the new regime often permits planners much discretion. State governments, for example, are unlikely to mandate training for all welfare recipients (Berger, Black, and Smith 2000; Dehejia forthcoming). Rather, some will be assigned to training, while others will receive standard benefits. Thus, there is a mixing problem (Manski 1997) in that the data reveal the outcomes under mandated treatments, but our interest is in learning what would happen if treatments are mixed across the population.

In light of the mixing problem, this paper examines what experimental evaluations reveal to planners faced with designing programs to meet performance standards. Abstracting from the well-known critiques of randomized experiments, I maintain the assumption that the evaluation identifies the average effect of the treatment. In practice, of course, there is often considerable disagreement about whether these data actually reveal the distribution of outcomes under mandatory treatment policies. There are many well-studied and important critiques of the validity of randomized social experiments.[1] The mixing problem, however, has received far less attention in the literature.[2] This fundamental identification problem, which arises even with valid experiments, may limit the planner's ability to assess the efficacy of a program.

1. See Campbell and Stanley (1966), Hausman and Wise (1985), Manski and Garfinkel (1992), Heckman and Smith (1995), and Manski (1996) for general critiques of the experimental methodology. Wiseman (1991) and Greenberg and Wiseman (1992) critically examine the MDRC demonstrations.

2. Manski (1997) and Dehejia (forthcoming) formally model the mixing problem under different assumptions.

The data alone cannot reveal the distribution of outcomes if the treatment assignment process may be mixed.

I develop this paper in two stages. In Sections II and III, I formalize the evaluation problem, making general and somewhat abstract points. In Sections IV–VI, I illustrate the methods by considering the specifics involved in designing programs to meet the new federal welfare-to-work standards.

Focusing on binary outcomes and treatments, Section II formalizes the basic concept of a performance standard and considers what experimental data reveal about the outcome of interest. Planners are assumed to compare expected outcomes with the standard, a threshold separating acceptable from unacceptable outcomes. While performance evaluation may be appealing in principle, the mixing problem creates ambiguity. An evaluator confronted with the mixing problem will necessarily combine the available data with assumptions on treatments and outcomes. In this setting, the performance of a program may be deemed acceptable, unacceptable, or indeterminate (Manski, Newman, and Pepper 2002).

Beginning with the same formal setup as in Manski (1997), Section III considers new constrained optimization models that might be used to address the mixing problem. These nonparametric models, which are consistent with commonplace theories of the selection process, bound but do not identify the distribution of interest. Still, the nonparametric bounds may be informative for performance evaluations.

After describing the methods in some generality, I then turn to the specific problem faced by welfare reformers in the new regime. Section IV considers what can be inferred from four employment focused welfare-to-work experiments conducted by the Manpower Demonstration Research Corporation (MDRC). Employment and job-readiness programs like those evaluated by the MDRC were used, in part, to motivate the federal reform,[3] have been used to inform planners in the new regime (see, for example, Michalopoulos, Schwartz and Adams-Ciardullo 2000), and are a key component of every state program.

After describing the four experiments, I evaluate what these data reveal about whether the program can meet the federal standard. In particular, I compare the estimated employment probability two years after the treatment is assigned to the long-run federal performance standard that half of single parent recipients participate in the labor force. In the absence of assumptions, these experiments only allow us to draw limited inferences about the labor market outcomes of welfare recipients. For the most part, the data do not reveal whether the federal standards will be met if a training program is adopted. In one case, however, these no-assumption estimates suggest that the program cannot meet the long-run federal standards. Under an outcome optimization model, the estimated bounds suggest that at least one-quarter and at most three-quarters of the caseload will participate in the labor force two years after the program is implemented. Under certain constrained optimization models, less than half of the welfare recipients can be expected to enter the labor market two years after the initial treatment.

3. The Riverside program in California, for example, has been used to motivate work first reform at the state and national levels (Hotz, Imbens, and Klerman 2000). While most demonstration programs, including the MDRC programs evaluated in Section IV, show modest employment effects, the Riverside program increased the employment rate from 35.3 percent for the control group to 49 percent for the treatment group.

Section V evaluates the sensitivity of the results to variation in parameters. With well-known concerns over the validity of welfare-to-work experiments, there will invariably be questions about the credibility of the empirical illustration. By exploring the sensitivity of the evaluation to variation in the performance threshold and observed outcomes, this analysis effectively defines the range of results that lead to conclusive findings about the efficacy of a program. These results can be used to evaluate alternative programs and arbitrary threats to validity.

Section VI concludes by considering what welfare-to-work demonstrations reveal to welfare reformers. Two substantive findings emerge. First, some programs cannot possibly meet the federal labor force participation requirement. Second, others may meet the requirement if there is enough heterogeneity in treatment response and if state/local officials act in way that optimizes the labor force participation probability.

II. Performance Standards: What Experimental Data Reveal

Consider the problem of evaluating whether a social program will meet a performance standard. Social programs determine the set of available treatments and influence the particular treatment that people receive. Formally, assume that for a given set of two mutually exclusive and exhaustive treatments an assignment policy, m, determines which treatment each person receives. Each person then realizes a binary outcome-of-interest that may depend on the treatment. Given the selection policy m, let z_m be the realized treatment and y_m be the realized outcome. The objective of the evaluation is to determine whether the program meets the performance standard, c. The usual practice is to evaluate programs in terms of their mean outcomes. Then, the program is acceptable if $P[y_m = 1] \geq c$.[4]

This program evaluation is straightforward if the outcome under the assignment policy m, y_m, is observed. Randomized experiments, however, may not reveal this outcome. Classically designed and properly implemented social experiments identify the distribution of outcomes that would occur if all individuals are given the same treatment. These data cannot reveal the distribution of outcomes that will occur when treatment assignment may vary across the population.[5]

To see this, assume each member k of a population K receives one of two feasible treatments and realizes a binary outcome-of-interest that may depend on this treatment. Let $y(1)$ and $y(0)$ denote these binary outcomes. In the welfare-to-work experiments, for example, $y(1)$ might measure the employment status that would occur were a welfare recipient to have been assigned to training and $y(0)$ the outcome that

4. To simplify the exposition, variables used to condition on observed characteristics of individual respondents are implicit. Here, covariates are used to define subgroups of interest, not to control for ''spurious effects.'' With sufficient prior information on the response functions or the selection policy, m, covariates might be included parametrically. Otherwise, nonparametric specifications are appropriate.

5. Alternative modes of data collection might yield information beyond what can be inferred from the classical experiments. One possibility would be a set of experiments where the randomization is at the site level rather than at the individual level. Here a treatment would be a set of rules under which case managers would operate, rather than a mandate. An experiment of this type would reveal how case managers would make treatment assignments. In principle, it would directly reveal the outcome of policy m.

would occur were she to receive standard benefits. Let z denote the actual treatment received, where $z = 1$ if assigned to treatment 1, and 0 otherwise. What do the experimental data reveal? For those who were assigned to treatment 1 ($z = 1$) the outcome indicator $y(1)$ is observed but $y(0)$ is latent, while for those who received treatment zero ($z = 0$) the outcome $y(0)$ is observed but $y(1)$ is latent. Thus, the data reveal the outcome probability for those who were assigned to treatment one, $P[y(1) = 1|z = 1]$, and for those who received treatment zero, $P[y(0) = 1|z = 0]$.

In social experiments, the actual treatment received is randomly assigned so that the treatment, z, is statistically independent of the outcome indicators, $y(1)$ and $y(0)$. Thus, data from valid social experiments identify the outcome probability if the population of interest were assigned to treatment one, $P[y(1) = 1]$, or if instead everyone were assigned to treatment zero, $P[y(0) = 1]$.

In practice, social programs may not mandate a single treatment to all recipients. Rather, the selection policy, m, may assign some recipients to receive treatment one, and others to receive treatment zero. In this case, the outcomes that would occur under the counterfactual social program with assignment policy m are not observable. The experiment cannot identify the probability a recipient receives a particular treatment nor the outcome probabilities among those persons who will receive that treatment.[6]

Thus, even abstracting from the well-known concerns about the validity of experimental evaluations, program evaluators will inevitably find it difficult to determine what constitutes an acceptable program. The experimental data alone cannot be used to implement the performance evaluation. Rather, inferences will depend critically on both the data and the prior information the evaluator can bring to bear. Almost no attention, however, has been given to resolving the ambiguity created by the mixing problem.

In fact, there is no unique resolution to this fundamental identification problem. If data on the outcome of interest are combined with sufficiently strong assumptions, the outcome probability under program m may be identified, implying a definitive performance evaluation. In practice, the most common assumption is that under the new regime all persons will receive a single treatment. Parametric latent variable models describing how treatments are selected and outcomes determined may also identify the outcome probability. Dehejia (forthcoming), for example, formalizes and estimates a parametric Bayesian model to evaluate welfare-to-work experiments. While his model identifies the labor market outcomes of interest, it rests on admittedly strong assumptions.

A social planner, concerned about the credibility of her findings to policymakers and the public, might be inclined to impose more conservative assumptions. Indeed, starting with the basic setup in Manski (1997), I evaluate what can be learned about the outcome distribution under policy m given weak assumptions on the process determining treatment selection. The result is a bound on $P[y_m = 1]$.

If the data and assumptions only suffice to bound $P[y_m = 1]$, the conventional idea of program evaluation using a single threshold to separate acceptable from unac-

6. Notice that the realized outcome under assignment policy m, y_m, is a probability mixture of $y(1)$ and $y(0)$. In particular, $y_m \equiv y(1)z_m + y(0)(1 - z_m)$. The distribution of realized outcomes is $P[y_m = 1] = P[y(1) = 1|z_m = 1]\,P[z_m = 1] + P[y(0) = 1|z_m = 0]P[z_m = 0]$.

ceptable outcomes needs revision (Manski, Newman, and Pepper 2002). Suppose it is known that $d_0 \leq P[y_m = 1] \leq d_1$, for known constants d_0 and d_1. Then the planner may conclude that for a given set of treatments, 0 and 1,

(1) Policy m is acceptable if $d_0 \geq c$ and unacceptable if $d_1 < c$.

Otherwise, the performance of the program is indeterminate.

III. Identification Under Alternative Assumptions

To formalize the identification problem, it is useful to first explore how selection policies might affect outcomes. Treatment only affects some individuals. In particular, a fraction $P[y(1) = 1 \cap y(0) = 0]$ of the population ''benefits'' from treatment one, while a fraction $P[y(1) = 0 \cap y(0) = 1]$ ''benefits'' from treatment zero. Regardless of the treatment assignment policy the failure rate will at least equal $P[y(1) = 0 \cap y(0) = 0]$ and the success rate must at least equal $P[y(1) = 1 \cap y(0) = 1]$. Thus, the joint distribution of outcome indicators, $y(1)$ and $y(0)$, implies that

(2) $P[y(1) = 1 \cap y(0) = 1] \leq P[y_m = 1] \leq 1 - P[y(1) = 0 \cap y(0) = 0]$.

Notice that the width of the bound in Equation 2 equals the fraction of individuals affected by the selection policy, $P[y(1) = 1 \cap y(0) = 0] + P[y(1) = 0 \cap y(0) = 1]$. In the absence of assumptions on the assignment policy, m, the joint distribution of $y(1)$ and $y(0)$ only identify the labor force participation probability if treatments have no effect on outcomes. Otherwise, the precise location of the outcome probability depends on how the planner assigns treatments among the affected populations.[7]

Since the data do not reveal the joint distribution of the labor force participation indicators, $y(1)$ and $y(0)$, the bounds in Equation 2 are not identified. This section formalizes what the demonstrations combined with prior assumptions on the treatment selection policy reveal about the outcome. Section IIIA reviews Manski's (1997) finding about what can be learned in the absence of assumptions. Section IIIB formalizes the implications of three new nonparametric restrictions on the selection process.

A. No-Assumption Bound

A logical first step is to examine what these data reveal in the absence of assumptions. In fact, the observed marginal distributions imply informative restrictions on the joint distribution in Equation 2. The ''no-assumption'' result is that knowledge of the outcome probabilities under homogeneous treatment policies yield a one-sided

7. In this paper, I assume the existence of a particular assignment rule, m, and the resulting realized outcome, y_m. The various bounds reveal what is known about the outcome probability, $P[y_m = 1]$, in different informational settings. Alternatively, one might consider the related normative question of how a planner should optimally assign treatments. In this paradigm, the nonparametric bounds reveal the range of $P[y_m = 1]$ over different possible assignment rules, m. For additional details, see Manski (2000) and Pepper (2002).

bound on the outcome probability under policy m. Formally, using the Frechet (1951) bounds,[8] Manski (1997, Proposition 1) shows that

$$(3)\quad \text{Max}\{0, P[y(1) = 1] + P[y(0) = 1] - 1\} \leq P[y_m = 1] \leq \text{Min}\{1, P[y(1) = 1] + P[y(0) = 1]\}.$$

There are two sources of uncertainty reflected in these bounds. First, as in Equation 2, the realized outcome probability depends on the unknown assignment rule, m. The upper (lower) bound, for example, is only realized if all who benefit from a particular treatment are assigned to that treatment (the alternative treatment). Second, additional uncertainty is introduced in that the data cannot reveal what fraction of the caseload is influenced by the treatment selection process.

Given the evaluation criteria in Equation 1, the performance under policy m will be unacceptable if the sum of the two outcome probabilities is less than the performance standard c, and acceptable if this sum minus one is greater than c. Otherwise, the performance is indeterminate.

B. Alternative Assumptions

In absence of additional data, the only way to resolve ambiguous findings is to impose assumptions. Here, I examine the implications of two easily understood and commonly suggested restrictions on the assignment process. The first model assumes administrators assign treatments to maximize the outcome probability. The second model restricts the fraction of recipients assigned to a particular treatment but makes no assumption about the selection rule.

1. Outcome Optimization

The outcome optimization model formalizes the assumption that decisions under uncertainty are made to maximize expected outcomes. Let w denote all relevant characteristics observed by the planner at the time of assignment. Assume that the planner has rational expectations, so that the probabilities $P[y(1) = 1|w]$ and $P[y(0) = 1|w]$ are known, and selects the treatment, z_m, that maximizes the outcome probability given these expectations.[9] Unable to distinguish between persons with the same covariates, w, the planner cannot implement assignment rules that systematically differentiate among these persons. Thus, given the observed characteristics w,

$$z_m = 1 \text{ if } P[y(1) = 1|w] > P[y(0) = 1|w]$$
$$\text{and } z_m = 0 \text{ if } P[y(0) = 1|w] \geq P[y(1) = 1|w].$$

Intuitively, under this assignment rule administrators can do no worse in terms of maximizing the outcome probability than what would have occurred if all recipi-

8. Heckman, Smith, and Clements (1997) and Heckman and Smith (1998) also use the Frechet bounds to show that experimental data cannot reveal the joint distribution of outcomes. They do not, however, evaluate the outcome distribution given the mixing problem.

9. This generalizes the model developed in Manski (1997, Proposition 6), which assumes planners know the response functions, $y(1)$ and $y(0)$. Dehejia (forthcoming) also examines a version of the outcome optimization model.

ents were assigned to a mandated treatment policy, and no better than the no-assumption upper bound in Equation 3. The resulting outcome distribution under this optimization model will depend upon the information known by the planner, w, as well as the fraction of the caseload affected by treatment. If the planner only observes the experimental results, then all persons would be assigned to a single treatment. In particular, $P[y_m = 1] = \max(P[y(1) = 1], P[y(0) = 1])$. If instead, the planner observes the response functions for each individual, then $y_m = \max[y(1), y(0)]$ and the treatment selection policy will maximize the outcome probability. In this case, $P[y_m = 1] = 1 - P[y(1) = 0 \cap y(0) = 0]$. Formally, the Frechet (1951) bounds imply the sharp restriction that

$$(4)\quad \max\{P[y(1) = 1], P[y(0) = 1]\} \leq P[y_m = 1] \leq \min\{P[y(1) = 1] + P[y(0) = 1], 1\}.$$

While the upper bound in Equation 4 coincides with the no-assumption upper bound in Equation 3, the lower bound is informative.

2. *Budget Constraint Model*

The budget constraint model formalizes the assumption that planners face resource and political constraints that may limit their ability to assign certain treatments. Welfare-to-work training programs, for instance, can be costly. Arguably, planners operate under a constraint limiting the fraction of recipients assigned to training.[10] Formally, suppose that no more than a certain known fraction, p, of individuals receive treatment one so that $P[z_m = 1] \leq p$.[11] The data and this restriction imply sharp bounds on the outcome distribution. In particular, I show in Appendix 1

> Proposition 1: For Bernoulli random variables $y(1)$ and $y(0)$, let $P[y(1) = 1]$ and $P[y(0) = 1]$ be known. Assume that $P[z_m = 1] \leq p$. Then,

$$(5)\quad \max\{0, P[y(0) = 1] + P[y(1) = 1] - 1, P[y(0) = 1] - p\} \leq P[y_m = 1] \leq \min\{1, P[y(0) = 1] + P[y(1) = 1], P[y(0) = 1] + p\}.$$

Notice that as the fraction who can receive treatment approaches one, the bounds converge to the no-assumption bound in Equation 3. In this case, the constraint is nonbinding. As the fraction approaches zero, however, the bounds center around the outcome that would be observed if all recipients receive treatment zero, $P[y(0) = 1]$. Restricting the fraction of recipients who can be trained narrows the bounds.

3. *Constrained Optimization Model*

In some cases, both the budget constraint and outcome optimization models may apply. That is, the objective of the planner might be to maximize the outcome proba-

10. Of course the costs of training may vary across participants, in which case, the budget constraint might not operate in the simple fashion described above.

11. Manski (1997, Proposition 7) evaluates the identifying power of a strict constraint specifying that a certain fraction of the caseload, p, receives treatment 1, and the remaining recipients, $1 - p$, receive treatment zero.

bility with a constraint limiting the fraction of recipients assigned to a particular treatment. Under the outcome optimization model, administrators assign treatments to those who benefit. Given this selection rule, the constraint is binding if the upper-bound fraction assigned to treatment one, p, falls below the proportion who benefit from this treatment.

Intuitively, under this constrained optimization model planners can do no worse than what would have occurred if all recipients were assigned to receive treatment zero and no better than the upper bound under the budget constraint model formalized in Proposition 1. In fact, I show in Appendix 1

Proposition 2: For Bernoulli random variables $y(1)$ and $y(0)$, let $P[y(1) = 1]$ and $P[y(0) = 1]$ be known. Assume that $P[z_m = 1] \leq p$ and that planners optimize outcomes. Then,

$$(6)\quad \max\{P[y(0) = 1], P[y(1) = 1]\, p + P[y(0) = 1]\,(1 - p)\} \leq P[y_m = 1] \leq \min\{1, P[y(0) = 1] + P[y(1) = 1], P[y(0) = 1] + p\}.$$

While the upper bound in Equation 6 coincides with the budget constraint upper bound in Equation 5, the lower bound is informative.

IV. Application: What Do Welfare-to-Work Experiments Reveal to Welfare Reformers?

Under the new welfare system, state governments must design and implement programs that both provide assistance and encourage work. In the short run, 25 percent of the caseload is required to work and in the long-run the employment probability must exceed 50 percent. Using data from four experiments conducted in the mid-1980s, I examine what welfare-to-work demonstrations reveal about outcomes when treatments may be mixed across the caseload. Section IVA describes the MDRC experiments. Section IVB explores the notion that the employment rate under policy m must lie between the observed marginal distributions. Finally, Section IVC presents the estimated bounds under the alternative models described above.

A. MDRC Demonstrations

This analysis exploits four well-known work and training demonstrations conducted in the 1980s by the Manpower Demonstration Research Corporation (MDRC): the Arkansas WORK Program, the Baltimore Options Program, the San Diego Saturation Work Initiative Model (SWIM), and the Virginia Employment Services Program (ESP). Table 1 provides summary information on the MDRC programs. To evaluate these experimental programs, the MDRC selected samples of size 1,127; 2,757; 3,211; and 3,150, in Arkansas, Baltimore, San Diego, and Virginia, respectively. For each program, welfare recipients were randomly assigned to either participate in a basic training activity, or to receive the standard benefits. For each respondent, the data reveal the treatment received—training or standard benefits—and numerous labor market and welfare participation outcome measures. In this paper, the outcome

Table 1
Selected Characteristics of the MDRC Welfare-to-Work Experiments

	Arkansas	Baltimore	San Diego	Virginia
Start date	1983	1982	1985	1983
Major program	Sequence of job search, unpaid work	Choice of job search, unpaid work, education	Sequence of job search, unpaid work, education	Sequence of job search, unpaid work
Age of children	3 or older	6 or older	6 or older	6 or older
Length of followup	36 months	36 months	27 months	33 months
Fraction complying with training	38.0%	45.0%	64.4%	58.3%
Net cost per person trained[a]	$118	$953	$919	$430
Observed outcomes	Employment, earnings, welfare receipt and payments	Employment, earnings, welfare receipt and payments	Employment, earnings, welfare receipt and payments	Employment, earnings, welfare receipt and payments
Sample size	1,127	2,757	3,211	3,150

Source: Gueron and Pauly (1991 Table 3.1) and Friedlander and Burtless (1995 Table 3.1).
a. Net costs are computed as the average costs per enrolled recipient minus the average costs for recipients in the control group. The net costs estimates include all expenditures by the operating agency as well as services provided by other agencies that were considered part of the treatment program.

variable of interest is whether or not the respondent participated in the labor force two years after treatment.

These MDRC experiments appear particularly well-suited for evaluating the types of welfare and training programs that might be adopted under TANF. Each of these evaluations were broad coverage programs with at least two years of follow-up data. All single parent families receiving AFDC and whose children were at least six years of age were mandated to participate. Similar to state TANF programs, noncompliance with the assigned treatment led to sanctions or possible expulsion from the program.[12]

Finally, the MDRC training programs stressed employment rather than human capital development. Educational activities were only offered in limited cases. Since the PRWORA stresses immediate employment, basic training programs like those evaluated by the MDRC are a defining feature of every state welfare program. In fact, the MDRC experiments and other similar job-readiness demonstration programs have been used to inform state and federal policy makers. Michalopoulos, Schwartz, and Adams-Ciardullo (2000), for example, evaluate 20 work-first demonstration programs, including SWIM, in a report prepared for the United States Department of Health and Human Services on the potential effectiveness of different welfare to work programs in the new regime.

While these demonstrations are generally recognized as well designed and implemented social experiments (see, for example, Gueron and Pauly 1991, Wiseman 1991, Greenberg and Wiseman 1992, and Friedlander and Burtless 1995), there are shortcomings as well. Concerns about external validity are particularly germane:[13] Both the economy and the welfare system have undergone major changes since the mid-1980s. Time limits, for example, did not exist prior to the reform.

In this section, I abstract from these concerns by assuming that the MDRC demonstrations identify the effects of the various job training programs. That is, up to sampling error, the data are assumed to reveal the labor force participation probability if all recipients are assigned to training, $P[y(1) = 1]$, and if all recipients are given the standard benefits, $P[y(0) = 1]$. In Section V, I evaluate the sensitivity of these results to variation in the parameters.

Table 2 displays the estimated employment probability for the treatment and control groups, along with the corresponding 90 percent bootstrapped confidence intervals.[14] In each case, the results suggest that job training programs slightly increase

12. All sample recipients assigned to training are assumed to experience a treatment. Most, in fact, received training. As in the current regime, those that did not comply either left AFDC or were sanctioned. Sanctions, which eliminated the parent's portion of the grant for between three and six months, were applied in fewer than 10 percent of the cases (Friedlander and Burtless 1995, Table 3–1). Under this intention to treat framework, all respondents comply with their assignment. If one wants to isolate the effects of the training program, as opposed to being assigned to training, noncompliance and substitution may become central concerns (see, for example, Bloom 1984 and Manski 1996).

13. See, for example, Hotz, Imbens, and Mortimer (1998) and Hotz, Imbens, and Klerman (2000), who examine what can be learned from the MDRC experiments when the composition of the caseload as well as the mix of programs may differ from those in place during the demonstration.

14. I apply the percentile bootstrap method. In particular, the bootstrap sampling distribution of an estimate of the upper bound is its sampling distribution under the assumption that the unknown population distribution equals the empirical distribution of these variables in the sample of respondents. Next to each upper (lower) bound estimate I report the 0.95 (0.05) quantile of its bootstrap sampling distribution.

Table 2
Estimated Probability of Employment Eight Quarters After Treatment

	Control Group: $P[y(0) = 1]$	Treatment Group: $P[y(1) = 1]$
Arkansas	20.5%	23.8%
	(0.175, 0.230)	(0.213, 0.265)
Baltimore	37.7%	38.8%
	(0.255, 0.392)	(0.373, 0.412)
San Diego	28.5%	35.0%
	(0.265, 0.302)	(0.331, 0.370)
Virginia	33.9%	39.0%
	(0.315, 0.362)	(0.373, 0.408)

Note: Bootstrapped 90 percent confidence intervals are in parentheses.

the probability of employment. Under the Virginia Employment Service Program (ESP), for instance, the employment probability if all welfare recipients receive training is estimated to be 39.0 percent. If instead, all welfare recipients receive the standard benefits, 33.9 percent would be working after two years. Thus, the ESP increases the probability of employment by 0.051.

B. Ordered Outcomes

Given the mixing problem, one might speculate that the employment probability under the new regime will necessarily lie between the outcomes under the mandated training and standard benefit treatments, with the precise location depending on what fraction of participants are assigned to training. This hypothesis is true if being assigned to training never reduces the likelihood of participating in the labor force. After all, under this ordered outcomes assumption, the planner can do no better in terms of maximizing the employment probability than assigning everyone to training and no worse than assigning everyone standard benefits. Thus, if outcomes are ordered, the estimates displayed in Table 2 imply that the training programs cannot achieve the long-run employment standards of the federal reform.

If instead the effects of treatment are heterogenous, with some fraction unaffected, some fraction employed only if assigned to training, and some fraction employed only if given standard benefits, the ordered outcomes bounds displayed in Table 2 do not apply. Planners can do better than assigning everyone to training and can do worse than assigning everyone to receive the standard benefits.

Arguably, in fact, the effects of treatment are heterogenous. Basic skills and job search programs are unlikely to benefit the entire welfare caseload that includes individuals with a broad range of skills, backgrounds, and challenges (Pavetti et al. 1997). Programs that enable recipients with low levels of human capital to transition into the labor force may not benefit those with different skills or impediments.

Table 3
Estimated No Assumption Bounds on the Probability of Employment Eight Quarters After Treatment Under Assignment Policy m, $P[y_m = 1]$

	Lower Bound	Upper Bound
Arkansas	0.0%	44.2% (0.474)
Baltimore	0.0%	76.5% (0.796)
San Diego	0.0%	63.5% (0.661)
Virginia	0.0%	72.9% (0.759)

Note: The 0.95 quantiles of bootstrapped upper bound are in parentheses.

Rather, for some fraction of the caseload, the time and resources devoted to training might otherwise be used more effectively.

In fact, there is some empirical support for this proposition. A nontrivial fraction of recipients assigned to training incur the costs of noncompliance (Friedlander and Hamilton 1993; Pavetti et al. 1997). Furthermore, the data reveal negative effects for certain observed subgroups. Consider, for instance, respondents who did not work in the quarter prior to random assignment but were employed in the quarter of assignment. Training for this subgroup lowers the employment probability by 11 points under the Arkansas WORK program, by four points under the Baltimore Saturation program, by 19 points under the San Diego SWIM program, and by three points under the Virginia ESP. Seemingly, training reduces the employment probability of respondents who appear to have relatively strong attachments to the labor market prior to the treatment.[15]

C. Estimated Bounds

Table 3 presents the no-assumption estimates for each of the four programs. Notice that these bounds are only informative on one side. In the absence of data, we know that the employment probability lies between zero and one. For each of the four programs, the data narrow the upper bound while the lower bound remains at zero.

Still, these bounds are informative. We learn, for instance, that regardless of the assignment policy, the employment probability under the Arkansas WORK program

15. The data cannot rule out the possibility that the ordered outcome assumption applies to respondents within observed subgroups. This restriction, however, seems tenuous. Since the effect of training is heterogenous across observed subgroups it seems plausible that there is unobserved heterogeneity as well.

will not meet the long-run labor force participation threshold to assure full TANF funding. At most, just more than 44 percent of the caseload will be participating in the labor force two years after the program is implemented. In contrast, the data are inconclusive about the Baltimore, San Diego, and Virginia programs. In the absence of assumptions, the data cannot reveal whether or not adopting these programs will achieve the long-run employment standards.

Suppose, however, that planners assign training to optimize the employment probability. Under the outcome optimization model, the planner can do no worse than assign everyone to training. Thus, the estimates in Table 2 suggest that under the Baltimore, San Diego, and Virginia programs, at least one-third of the caseload will work. However, even in this best-case model there remains much uncertainty. Consider, for instance, Virginia's ESP program. If planners combine this program with an outcome optimization assignment rule, the estimated bounds imply that at least 39.0 percent and at most 72.9 percent of welfare recipients will be employed after two years. Where the realized labor force participation probability will lie depends upon prior information, w, and the association between the latent labor force participation indicators, $y(1)$ and $y(0)$. If these outcomes have a strong positive association, the realized probability will lie closer to the lower bound regardless of the planner's prior information. In contrast, if the association is strongly negative, the realized probability will approach the upper bound if the planner has sufficient prior information on the response functions.

Alternatively, planners might face a constraint on the number of recipients who can be trained. After all, even the modest training programs evaluated by the MDRC are costly. Table 4 displays estimated bounds under the constraint that no more than 10 percent or 25 percent of the caseload will be assigned to training. If restricted to training one-tenth of the caseload, the Baltimore Options Program will meet the short-run federal labor force participation standard of 25 percent, while the San Diego SWIM and Virginia ESP programs might meet this standard. In all cases, however, employment outcomes under these programs fall short of the long-run performance standard. In contrast, if planners train less than one-fourth of the caseload, the estimated bounds are too wide to infer whether or not the programs will meet any of the federal labor force participation standards. With increased flexibility planners might either improve or degrade the decision making process.

Arguably, both the budget constraint and outcome optimization models apply. Table 4 also displays the estimated constrained optimization bounds. Under the Baltimore, San Diego, and Virginia programs more than one-quarter of the caseload will be working two years after the treatment is assigned. The short-run employment standards will be met. Whether the long-run requirements are achieved depends on both the constraint and the association between the outcomes. As before, if only 10 percent of the caseload can be assigned to training, the programs cannot meet the long-run employment standard. If instead one-fourth of the caseload can be assigned to training, the upper bound exceeds the 50 percent benchmark in all cases. Thus, if planners have sufficiently detailed prior information and if outcomes are negatively associated, the long-run federal benchmark may be achieved. If the outcomes have a strong positive association, however, policy m cannot meet the federal benchmark.

Table 4
Estimated Bounds on the Probability of Employment Eight Quarters After Treatment Under the No-Assumption and Optimization Budget Constraint Models

	Lower Bound		
Selection Model	No-Assumption	Optimization	Upper Bound
No more than 10 percent trained ($p = 0.10$)			
Arkansas	10.5%	20.8%	30.5%
	(0.075)	(0.179)	(0.334)
Baltimore	27.7%	37.8%	47.7%
	(0.256)	(0.267)	(0.499)
San Diego	18.5%	29.2%	38.5%
	(0.165)	(0.272)	(0.402)
Virginia	23.9%	34.4%	43.9%
	(0.215)	(0.321)	(0.462)
No more than 25 percent trained ($p = 0.$			
Arkansas	0.0%	21.3%	44.2%
	(0.000)	(0.185)	(0.474)
Baltimore	12.7%	38.0%	62.7%
	(0.106)	(0.285)	(0.649)
San Diego	3.5%	30.1%	53.5%
	(0.015)	(0.282)	(0.552)
Virginia	8.9%	35.2%	58.9%
	(0.065)	(0.330)	(0.612)

Note: The 0.95 quantiles of bootstrapped upper bound are in parentheses below the upper bound estimates and the 0.05 quantiles of the bootstrapped lower bound are in parentheses below the lower bound estimates.

V. Sensitivity Analysis

Although the MDRC experiments and others like them have been used to inform policy makers in the new regime, there is likely to be concern with the validity of the demonstrations. Arguably, experiments conducted prior to the reform may not be relevant to the problems faced by planners today. Both the low-skilled labor market and the welfare system have changed over this period. In addition to concerns about external validity, the basic job training programs evaluated by the MDRC may not be relevant: Planners may be using different training programs and incentive schemes than those explored by the MDRC. Finally, there remains much uncertainty about how the performance standards will be implemented in practice. State governments are given some power to define both the numerator (work activities) and the denominator (the caseload and eligibility criteria) when computing labor force participation rates.

In this section, I evaluate the sensitivity of the performance evaluation to variation in known parameters. In particular, this analysis defines the set of results that lead to conclusive findings about the efficacy of a program. The information enables one to evaluate other welfare-to-work demonstration programs and, more generally, alternative social programs and performance standards. It also characterizes the degree to which concerns about validity might lead to biased performance evaluations.

Let P_{c0} and P_{c1} be conjectured values for the outcome probability under treatment zero and one, respectively. Then, given the no-assumption bounds in Equation 3 and the performance evaluation in Equation 1, define the switching thresholds for the performance standard as:

$$(7) \quad \max(P_{c0} + P_{c1} - 1, 0\} = c_A \quad \text{and} \quad \min(P_{c0} + P_{c1}, 1) = c_U,$$

where c_A measures the maximum performance threshold that would result in the performance being acceptable, and c_U measures the minimum threshold that would result in the performance being unacceptable.

Table 5 displays the switching thresholds given no prior information about mixing process or the outcome distribution. The first panel displays the maximum performance threshold that would result in the program being acceptable, c_A, and the second panel displays the minimum performance threshold that would result in the program being unacceptable, c_U. The shaded figures represent outcomes that would lead to unambiguous decisions given the long-run performance measure used in the recent welfare reforms, namely that at least half the caseload be employed. As revealed in Equation 7, the thresholds are linear in the marginal probabilities. Thus, in the absence of assumptions, acceptable programs are only found when the sum of the marginal probabilities is relatively high; unacceptable programs are found when this sum is relatively low.

Switching thresholds under alternative models can be defined similarly. For example, the same thresholds apply under the budget constraint model, except that when the constraint is binding the defining parameter is p, rather than the conjectured marginal probability, P_{c1}. Notice that binding constraints increase the chances of an unambiguous evaluation. By effectively restricting the planner's ability to assign treatments, constraints decrease the upper bound and increase the lower bound.

The switching thresholds under the optimization models also follow quite simply. Clearly, there is no effect on the unacceptable region—the upper-bound employment probability does not change. The model does, however, increase the lower bound so that maximum acceptable performance threshold increases, sometimes quite substantially. In particular, without a budget constraint, the threshold increases to the maximum of the two marginal probabilities. Thus, the threshold is greater than zero, and increasing with the maximum marginal employment probability. Acceptable plans under the existing welfare system occur once the employment rate under mandatory training (or cash assistance) exceeds 0.5.

These thresholds effectively define the requirements for unambiguous evaluations. Abstracting from sampling variability, the results in Table 5 can be used to evaluate any basic welfare to work experiment. Consider, for example, evaluating the Riverside demonstration program conducted in California in the late 1980s and early 1990s. This program, which is one of the most successful and widely cited welfare-

Table 5
Switching Thresholds for the Performance Standard Given Conjectured Values for the Marginal Probabilities Under the No-Assumption and Budget Constraint Models

P_{c0}

c_A

$\max\{P_{c1}, 1-p\}$		**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
	0.1	0	0	0	0	0	0	0	0	0
	0.2	0	0	0	0	0	0	0	0	0.1
	0.3	0	0	0	0	0	0	0	0.1	0.2
	0.4	0	0	0	0	0	0	0.1	0.2	0.3
	0.5	0	0	0	0	0	0.1	0.2	0.3	0.4
	0.6	0	0	0	0	0.1	0.2	0.3	0.4	0.5
	0.7	0	0	0	0.1	0.2	0.3	0.4	0.5	0.6
	0.8	0	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
	0.9	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8

c_U

$\min\{P_{c1}, p\}$		**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1
	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1	1
	0.4	0.5	0.6	0.7	0.8	0.9	1	1	1	1
	0.5	0.6	0.7	0.8	0.9	1	1	1	1	1
	0.6	0.7	0.8	0.9	1	1	1	1	1	1
	0.7	0.8	0.9	1	1	1	1	1	1	1
	0.8	0.9	1	1	1	1	1	1	1	1
	0.9	1	1	1	1	1	1	1	1	1

Note: P_{c0} and P_{c1} are conjectured values for the outcome probability under treatment zero and one, respectively. p denotes the maximum fraction of the caseload that can be assigned to treatment 1. c_A measures the maximum performance threshold that would result in the performance being acceptable, and c_U measures the minimum threshold that would result in the performance being unacceptable.

to-work training demonstrations, revealed employment rates for the treatment group of 49 percent and for the control group of 35 percent (Hotz, Imbens, and Klerman 2000). Thus, in the absence of assumptions about assignment policy m, the Riverside program will lead to an unacceptable evaluation for any performance standard greater than 0.86. Otherwise, the evaluation will be indeterminate. Under the optimization model, however, the Riverside program nearly leads to an acceptable evaluation in the current regime.

VI. Conclusion

Understanding the outcomes that can be expected from implementing a job training program is a central concern to academics, policymakers, and program administrators. What will the post welfare reform world look like? What fraction of the caseload will work? How will poverty and welfare spells change? What will happen to the teenage pregnancy rate? Almost no empirical evidence has been brought to bear on these and other important questions.

Under the new federal regulations, state and local governments must design and implement welfare programs that meet minimum labor force participation requirements. In many cases, the only available information on innovative programs comes from welfare-to-work demonstrations conducted since the War on Poverty. What do these experiments reveal to welfare reformers? Two general findings emerge. First, some programs cannot meet the federal standards. Second, other programs may meet the requirements if there is both sufficient heterogeneity in the treatment response, and planners optimize outcomes given full information about the latent response functions, or something reasonably close. While achieving this latter requirement depends upon the information and objective of the planners, the former requirement depends upon the fraction of the caseload affected by treatment.

Consider, for example, the Virginia Employment Services Program. In the absence of any restrictions to address the mixing problem, the employment probability under the ESP falls within [0, 0.729]. Thus, the experimental data do not reveal whether instituting the ESP will achieve the federal employment standards.

Prior information substantially narrows the no-assumption bound. If, for instance, administrators assign treatments to optimize the employment probability, the experiments imply that the program will achieve the short-run federal standards. At least 39 percent of the caseload will work. Even under the outcome optimization model, however, the data do not reveal whether outcomes under the ESP program can meet the long-run federal labor force participation standards. After all, the experimental data cannot reveal the fraction of the caseload affected by the assignment process. If instead, we assume that the planners are restricted to train 10 percent of the caseload, the labor force participation probability will lie between [0.239, 0.439]. Under this restriction, we learn that the caseload may meet the short-run federal standards but cannot achieve the long-run objectives.

While there may be doubts about the validity of the MDRC experiments, similar qualitative conclusions will be found from nearly every experimental evaluation conducted prior to the reform. This does not imply that reformers cannot succeed in meeting the performance threshold. An indeterminate finding does not mean that the performance will ultimately be unacceptable. Furthermore, it might be that external factors bias the validity of experiments. Time limits, state discretion in defining work-related activities, and other assistance programs (for example, transportation and childcare assistance) may all lead to higher employment rates than found in the prereform experiments. The switching thresholds displayed in Table 5, however, apply regardless. Valid experiments conducted under the current regime will only lead to unambiguous findings if these thresholds are met.

Appendix 1

Proofs of Propositions 1 and 2

In this appendix, I derive nonparametric bounds under restrictions on the fraction assigned to treatment one. Two cases are examined. First, I prove Proposition 1 which bounds the outcome distribution under the weak budget constraint model. Second, I prove Proposition 2 which bounds the outcome distribution under the constrained optimization model.

A. Proposition 1: Weak Budget Constraint Model

Suppose the weak constraint that $P[z_m = 1] \leq p$ applies. The constraint limits the planner's ability to assign treatment 1 to those affected by that treatment. Thus, the no-assumption bound on the outcome distribution in Equation 2 can be modified such that

$$\begin{aligned}(A1)\quad & P[y(1) = 1 \cap y(0) = 1] + \max(0, P[y(1) = 0 \cap y(0) = 1] - p) \\ & = \max(P[y(1) = 1 \cap y(0) = 1], P[y(0) = 1] - p\} \leq P[y_m = 1] \leq \\ & \quad 1 - P[y(1) = 0 \cap y(0) = 0] - \max(0, P[y(1) = 1 \cap y(0) = 0] - p) \\ & = P[y(0) = 1] + \min(P[y(1) = 1 \cap y(0) = 0], p\}.\end{aligned}$$

Notice that the lower bound on $P[y_m = 1]$ increases in $P[y(1) = 1 \cap y(0) = 1]$ whereas upper-bound increases in $P[y(1) = 1 \cap y(0) = 0]$. Thus, a sharp lower bound on $P[y_m = 1]$ is found by setting $P[y(1) = 1 \cap y(0) = 1] = \max\{0, P[y(0) = 1] + P[y(1) = 1] - 1\}$ and a sharp upper bound is found by setting $P[y(1) = 1 \cap y(0) = 0] = \min(P[y(1) = 1], P[y(0) = 0])$ (Frechet 1951).

B. Proposition 2: Outcome Optimization with a Weak Budget Constraint

The outcome optimization model assumes that planners optimize outcomes given the available information, w. Given this model, the resulting outcome distribution varies with both the information set, w, and the fraction of recipients affected by treatment.

Thus, the sharp upper bound can be found by evaluating the outcome distribution in the best case scenario when the observed covariates, w, identify response functions. In this case, the planner can implement the optimal decision rule, $y_m = \max(y(1), y(0))$ for each individual and the upper bound coincides with the upper bound found under the budget constraint model in Proposition 1.

Likewise, the sharp lower bound can be found by evaluating the outcome distribution in the worst case scenario where the information set, w, only reveals the covariates included in the experimental evaluation. Although the planner cannot implement assignment rules that systematically differentiate among persons with the same observed characteristics, w, she can randomly assign different treatments to persons with the same covariates. With random assignment, $P[y_m = 1 | z_m = t] = P[y(t) = 1]$.

Thus, subject to the constraint that $P[z_m = 1] \leq p$, the planner assigns treatments to maximize

(A2) $P[y_m = 1] = P[y(1) = 1]P[z_m = 1] + P[y(0) = 1]\,P[z_m = 0]$.

If $P[y(0) = 1] \geq P[y(1) = 1]$, the planner assigns treatment zero to all recipients. Otherwise, the planner randomly assigns heterogenous treatments, such that $P[z_m = 1] = p$.

References

Berger, Mark, Dan Black, and Jeffrey Smith. 2000. ''Evaluating Profiling as Means of Allocating Government Services.'' University of Western Ontario.

Bloom, Howard S. 1984. ''Accounting for No-Shows in Experimental Evaluation Design.'' *Evaluation Review* 8(2): 225–46.

Campbell, Donald, and Julian Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research*. Boston, Mass.: Houghton Mifflin.

Dehejia, Rajeev. Forthcoming. ''Program Evaluation as a Decision Problem,'' *Journal of Econometrics*.

Frechet, Maurice. 1951. ''Sur Les Tableaux de Correlation Donte les Marges Sont Donnèes.'' *Annals de Universitè de Lyon A* 3(14): 53–77.

Friedlander, Daniel, and Gary Burtless. 1995. *Five Years After: The Long Term Effects of Welfare-to-Work Programs*. New York: Russell Sage Foundation.

Friedlander, Daniel, and Gayle Hamilton. 1993. *The Saturation Work Model in San Diego: A Five-Year Follow-Up Study*. New York: Manpower Demonstration Research Corporation.

Greenberg, David, and Michael Wiseman. 1992. ''What Did the OBRA Demonstrations Do?'' In *Evaluating Welfare and Training Programs*, Charles F. Manski and Irwin Garfinkel, eds., 25–75. Cambridge, Mass.: Harvard University Press.

Gueron, Judith M., and Edward Pauly. 1991. *From Welfare to Work*. New York: Russell Sage Foundation.

Hausman, Jerry A., and David A. Wise, eds. 1985. *Social Experimentation*. Chicago: University of Chicago Press.

Heckman, James J., and Jeffrey Smith. 1995. ''Assessing the Case for Social Experiments.'' *Journal of Economic Perspectives* 9(2): 85–110.

———. 1998. ''Evaluating the Welfare State.'' In *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*, Strom, Steiner, ed., 241–318. Cambridge, England: Cambridge University Press.

Heckman, James J., Jeffrey Smith, and Nancy Clements. 1997. ''Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts.'' The *Review of Economic Studies* 64(4): 487–536.

Hotz, V. Joseph, Guido W. Imbens, and Jacob A. Klerman. 2000. ''The Long-Term Gains from GAIN: A Re-Analysis of the Impacts of the California GAIN Program.'' Cambridge, Mass.: NBER Working Paper, 8007.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 1998. ''Predicting the Efficacy of Future Training Programs Using Past Experiences.'' Los Angeles: University of California—Los Angeles Working Paper.

Manski, Charles F. 1996. ''Learning About Treatment Effects from Experiments with Random Assignment of Treatments,'' *Journal of Human Resources* 31(4), 707–33.

———. 1997. ''The Mixing Problem in Programme Evaluations.'' *The Review of Economic Studies* 64(4): 537–53.

———. 2000. ''Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice.'' *Journal of Econometrics* 95(2), 415–42.

Manski, Charles F., and Irwin Garfinkel, eds. 1992. *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press.

Manski, Charles F., John Newman, and John V. Pepper. 2002. ''Using Performance Standards to Evaluate Social Programs with Incomplete Outcome Data.'' *Evaluation Review* 26(4), 355–81.

Michalopoulos, Charles, Christine Schwartz, and Diana Adams-Ciardullo. 2000. ''National Evaluation of Welfare-to-Work Strategies: What Works Best for Whom: Impacts of 20 Welfare-to-Work Programs by Subgroup.'' Manpower Demonstration Research Corporation: New York, N.Y.

Pavetti, LaDonna, Krista Olson, Demetra Nightingale, Amy-Ellen Duke, and Julie Isaacs. 1997. *Welfare-to-Work Options for Families Facing Personal and Family Challenges: Rationale and Program Strategies*. Urban Institute Press: Washington, D.C.

Pepper, John V. 2002. ''To Train or Not To Train: Optimal Treatment Assignment Rules Under the New Welfare System.'' Charlottesville, Va.: University of Virginia, Department of Economics, Thomas Jefferson Center Discussion Paper.

Wiseman, Michael, ed. 1991. ''Research and Policy: A Symposium on the Family Support Act of 1988.'' *The Journal of Policy Analysis and Management* 10(4):588–666.