MDRC Working Papers on Research Methodology

Extending the Reach of Randomized Social Experiments: New Directions in Evaluations of American Welfare-to-Work and Employment Initiatives

James A. Riccio
Howard S. Bloom

Manpower Demonstration
Research Corporation

MDRC

Revised October 2001

This is part of a series of MDRC working papers that explore alternative methods of evaluating the implementation and impacts of social programs and policies.

An earlier version of this paper was prepared for a meeting of the Royal Statistical Society in London on July 4, 2000. The current version will be published in the Journal of the Royal Statistical Society"

Work on the paper was supported by a grant from The Rockefeller Foundation to further a U.S.-U.K. dialogue on evaluation research, a grant from the Pew Charitable Trusts to promote the development of new quantitative and qualitative evaluation research methodologies, and a grant from the Russell Sage Foundation to prepare a book on combining experimental and non-experimental methods for measuring the impacts of social programs.

Dissemination of MDRC publications is also supported by the following foundations that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: the Ford, Ewing Marion Kauffman, Ambrose Monell, Alcoa, George Gund, Grable, Starr, Anheuser-Busch, New York Times Company, Heinz Family, and Union Carbide Foundations; and the Open Society Institute.

The findings and conclusions presented here do not necessarily represent the official positions or policies of the funders.

The authors would like to thank Hans Bos, Lisa Gennetian, Virginia Knox, Charles Michalopoulos, and Pamela Morris for their insights on the issues discussed in this paper.

ABSTRACT

Random assignment experiments are widely used in the United States to test the effectiveness of new social interventions. This paper discusses several major welfare-to-work experiments, highlighting their evolution from simple "black box" tests of single interventions to multi-group designs used to compare alternative interventions or to isolate the effects of components of an intervention. The paper also discusses new efforts to combine experimental and non-experimental analyses in order to test underlying program theories and maximize the knowledge gained about the effectiveness of social programs. Researchers and policymakers in other countries may find this variety of approaches useful to consider as they debate an expanded role for social experiments.

1. Introduction

During the last three decades, social experiments have been used extensively in the United States to determine the effectiveness of social programs and have played an especially prominent role in the evaluation of welfare-to-work and employment programs (Friedlander, *et al,*. 1997; Greenberg and Shroder, 1997). This fact is a testament to the widely held belief that such experiments, which employ random assignment to allocate subjects to treatment and control groups, are the best way to produce unbiased estimates of program effects (e.g., Betsey, *et al.*, 1985; Boruch, 1997). It is also a testament to the growing recognition that in many settings it is possible to conduct random assignment on a large scale in a fair and ethical manner (Boruch, 1997; Gueron, 1999).

Random assignment to a new treatment is generally accepted as ethical when the effectiveness of the treatment is unknown, when subjects for the experiment are not being denied access to an opportunity to which they are legally entitled, and when resources are not sufficient to provide access for everyone who might qualify. In this situation, random assignment can be viewed as a fair way of allocating scarce resources, because it gives everyone involved the same chance to receive them.

However, random assignment is not always technically feasible, even when the political will to use it exists. In some cases, the nature of the intervention (for example, those focused on whole groups) makes it impractical to choose individual subjects on a random basis. In other cases, random assignment is too blunt an instrument for testing important program theories or hypotheses. Thus, although it has proven to be a powerful tool for assessing social policies, random assignment is often limited by technical and practical constraints.

The reality of these constraints poses a challenge for social researchers: how to work within and around them in order to maximize the value of social experiments for knowledge development. In particular, when the goal is to build a convincing body of knowledge in a social program area about "what works, and why," it is important to consider how the experimental method can be adapted, extended, and supplemented by other approaches.

This paper illustrates recent efforts to do just that with social experiments involving American welfare-to-work and employment programs conducted by the Manpower Demonstration Research Corporation (MDRC), a New York-based social policy research firm. Many of the issues discussed in the paper will be explored further in a book that is being written by MDRC for the Russell Sage Foundation.

During the past 25 years, MDRC has conducted 30 major random assignment experiments involving nearly 300,000 people (Gueron, 1999). The collection of experiments described here—many of which are still in progress or just beginning—reveals an ongoing evolution in analytical approaches. Implicit in these new directions is a view that random assignment cannot stand alone, and that researchers and policymakers must look beyond the "experimental *versus* non-experimental" debate toward productive ways of combining both approaches.

The paper begins by reviewing the use of a simple random assignment design to test the overall effects of a welfare-to-work program. It then highlights the advantages and limitations of this method and illustrates how some of the limitations can be overcome by a multi-group random assignment design, which has been used by several recent studies. The paper next discusses new efforts to test a broader range of hypotheses concerning factors that drive program performance. These efforts involve, in one example, pooling data from multiple experiments and applying multi-level modeling techniques to a comparison-of-sites analysis, and, in another case, using instrumental variables within the context of social experiments to test hypothesized causal relationships between a program's intermediate effects and its ultimate impacts. Finally, the paper discusses the mixing of cluster random assignment (random assignment of groups) and a quasi-experimental interrupted time-series design to study the effectiveness of a new employment initiative targeted on whole communities—specifically, public housing developments.

2. Basic Randomized Designs That Test Programs in their Entirety (Testing the "Black Box")

Simple randomized experiments to test social programs work much like medical clinical trials to test new drugs. Individuals are randomly assigned through a lottery-like process to either a program group, which is offered the new treatment, or a control group, which is not. This procedure ensures that the two groups are equivalent on all pre-random-assignment characteristics, both measured and unmeasured (assuming a sufficiently large sample) because each sample member has the same chance of being selected for the program. Thus, the subsequent outcomes of the control group accurately reflect what the program group's outcomes would have been in the absence of the program. Therefore, any observed differences between the two groups' future outcomes can be attributed to the program. For example, if 50 percent of the persons assigned to a welfare-to-work program found employment compared with only 40 percent of control group members, the 10 percentage point difference in their employment rates would represent the effect, or "impact," caused by the program.

Most experiments conducted on welfare-to-work or job-training programs test the program in its entirety instead of testing its separate components. Doing so is sometimes referred

to as testing a "black-box," not what is inside of it. An illustration of this approach is provided by MDRC's evaluation of California's Greater Avenues for Independence Program (GAIN).

GAIN was created by the California legislature in the mid-1980s to promote a more work-focused welfare system that balanced new opportunities for recipients with new obligations. The program was targeted on applicants for and recipients of Aid to Families with Dependent Children (AFDC), the national cash welfare system in the US at the time. The GAIN program offered a variety of work-preparation activities provided in specific sequences for different types of participants. Job-search assistance and basic education or English language instruction were the primary initial activities, followed, if appropriate, by vocational training, post-secondary education, or unpaid work experience in a public or not-for-profit organization. The program also provided financial assistance for child-care and other support services, plus assignment to a case manager. Welfare recipients who failed to participate in the program without "good cause" were to be "sanctioned" through a reduction in their welfare grant.

The GAIN model was the result of a political compromise among legislators, who debated how much the program should emphasize human capital development activities versus rapid employment, how much it should stress unpaid work experience, and whether participation in the program should be mandatory or voluntary. In the end, the new program stood in sharp contrast in scope, intensity, complexity, and expense to the job-search-only or job search/work experience programs of the early 1980s (Gueron and Pauly, 1991; Friedlander and Burtless, 1995). Given the many changes embodied in this new approach, lawmakers wanted to know, with confidence: "Does it make a difference"?

To answer this question, a randomized experiment was launched in six California counties (Alameda, Butte, Los Angeles, Riverside, San Diego, and Tulare) that reflected much of the state's diversity in welfare populations and local labor markets. Within each county, welfare applicants and recipients with school-age children were randomly assigned to the GAIN program or to a control group that was excluded from the program. Control group members could seek alternative services in the community on their own, but they were not required to do so. The core sample for the study included nearly 23,000 single parents (mostly mothers), about 22 percent of which were assigned to the control group. The employment and welfare experiences of the two groups, along with other outcomes, were then tracked for five years after each sample member's date of random assignment using administrative records. Based on this follow-up information, program impacts were estimated separately for each county, thereby providing six large-sample tests of the GAIN model (Riccio, *et al.*, 1994; Freedman, *et al*., 1996).

During the five-year follow-up period, GAIN produced statistically significant and substantively important impacts in each county. However, one county, Riverside, strongly outperformed the others by producing larger impacts on more variables across more subgroups of welfare recipients. At the opposite extreme, Los Angeles was distinguished by its lack of impacts on earnings. Overall, though, the six county tests suggested that the GAIN model was fairly "robust" and thus could produce positive effects when operated by different staff, for different types of people, in diverse settings.

Still, the variation in county performance was substantial and the randomized experiment was less helpful with respect to understanding what caused the variation than it was in documenting its existence. To address this question, a detailed study of the variation in county GAIN program practices was undertaken, drawing on a rich body of field observations in local GAIN offices and an extensive survey of their staff perceptions and practices. (These variations in practices reflected the fact that California's welfare system allowed county officials a certain amount of flexibility in how they implemented state policies.) In addition, differences across counties in the characteristics of sample members were controlled for statistically, and the variation in local labor market conditions was examined.

This analysis suggested that Riverside's superior performance derived from its *combination* of practices rather than any single practice. Key features of these practices included: a pervasive emphasis on moving recipients quickly into the labor market, backed up by staff efforts to locate job openings; a relatively balanced use of job search and education activities for recipients needing remedial education; a commitment to working with all enrollees, not just the most job-ready; close monitoring of recipients' participation in the program; and firm enforcement of the program's participation mandate (Riccio, *et al.*, 1994; Riccio and Hasenfeld, 1996; and Riccio and Orenstein, 1996).

Although the assessment of cross-county variation in program performance was carefully done, considerable uncertainty remained because it was not possible to isolate the effect of specific factors, given the large number of factors that varied across the small number of counties involved. Nevertheless, the analysis cast doubt on the labor market benefits of GAIN's huge investment in basic education because those counties that increased participation in this component the most (relative to the control group) experienced some of the smallest program impacts that were observed.

3. Multi-Program Group Designs that Compare Program Strategies and Components (Prying Open the "Black Box")

The policy debates in California over the structure of GAIN reflected many nationally relevant themes. Thus, it was not surprising that new federal welfare legislation enacted in 1988 (the Family Support Act) fostered welfare-to-work programs with many features similar to those of GAIN. In addition this legislation created an opportunity for conducting more rigorous tests of specific welfare-to-work strategies, not just whole programs.

Toward this end, the National Evaluation of Welfare-to-Work Strategies (NEWWS) sponsored by the U.S. Department of Health and Human Services, launched a series of social experiments in seven localities in six states. In four of these localities, policymakers agreed to conduct experiments with two different treatment groups plus a control group. This multi-program group design provided "head-to-head" experimental comparisons of alternative job preparation strategies (in three states) and alternative ways to structure program caseloads (in one state).

During the same period, the state of Minnesota launched an experiment that used a three-group randomized design to help isolate the effects of a financial incentive that was part of the

state's new welfare-to-work strategy—The Minnesota Family Investment Program (MFIP). To illustrate the potential and limitations of these multi-program group designs, key features and findings of the NEWWS and MFIP experiments are described briefly below.

- **Testing Human Capital Development Versus Labor Force Attachment Approaches (NEWWS)**

Randomized experiments in three NEWWS sites—Atlanta, Georgia; Grand Rapids, Michigan; and Riverside, California—tested two competing theories about how best to help individuals make a lasting transition from welfare to work: (1) by encouraging them to take a job (any job) quickly as a stepping stone toward long-term economic self-sufficiency, or (2) by inducing them to participate in education and training activities designed to increase their human capital and thereby improve their future economic prospects (see Hamilton, *et al.*, 1997).

To test these two theories, welfare applicants and recipients were randomly assigned to one of three groups at each site: (1) a "labor force attachment" group that emphasized rapid employment, (2) a "human capital development" group that emphasized education and training prior to employment, or (3) a control group that was not offered program services but was free to seek them from elsewhere in the community. The human capital strategy emphasized basic education as its initial activity for recipients with weak basic skills or limited prior education, and vocational training as its initial activity for recipients with a high school degree or its equivalent. In addition, the strategy provided job search assistance as a subsequent activity but did not emphasize it. In contrast, the labor force attachment strategy emphasized job search assistance as its first and primary activity; the strategy only provided education and training when its first component was not successful.

Findings from the study's implementation research indicated that local program staff delivered—and recipients in the different program groups heard—quite different messages about how best to move from welfare to work. Moreover, a cost analysis revealed that the human capital development programs were about twice as expensive as the labor force attachment programs. Thus, there was a substantial difference between the two program options tested.

The three-group randomized design used by these sites provided unbiased estimates of three types of impacts: (1) the *net* impacts of the labor force attachment strategy (from the difference in outcomes for the first program group and the control group), (2) the *net* impacts of the human capital development strategy (from the difference in outcomes for the second program group and the control group), plus (3) the *differential* impacts of the two strategies (from the difference in outcomes for the two program groups).

Interim NEWWS findings for the first two years after random assignment indicate that both program strategies had positive net impacts on total two-year earnings (Hamilton, et al., 1997). However, the impacts of the labor force attachment strategy tended to be larger than those of the human capital strategy. This may be due, in part, to the fact that a number of human capital group members were in school (and thus not employed) for a portion of the two-year follow-up period. At the present time, five years of NEWWS follow-up data have been collected but the

analysis of these data is not complete. Thus, it is not yet clear whether the short-term superiority of the labor force attachment approach will persist over time.

- **Testing Integrated Versus Specialized Case Management (NEWWS)**

A fourth NEWWS site—Columbus, Ohio—used a three-group randomized design to test alternative ways of structuring case management in welfare-to-work programs. Case managers in such programs must become employment specialists who perform a variety of tasks to help welfare recipients get and keep jobs. These tasks typically include, among others: assessing client employment barriers, guiding them toward specific program activities, arranging family support services, monitoring individual participation, and, when necessary, initiating financial sanctions for lack of program participation. Traditionally, these functions are kept separate from those of income maintenance workers, who must authorize and process welfare checks, determine program eligibility, and implement financial sanctions when notified to do so.

Proponents of separating the income maintenance and employment functions of welfare-to-work programs argue that doing so allows staff to specialize in a narrow range of functions which may call for different skills. This specialization could in turn, foster more efficient and higher-quality performance of each set of functions. On the other hand, critics of separating these functions argue that doing so slows the movement of welfare recipients into work-promoting activities and makes it tougher for state and local administrators to change the prevailing culture of the welfare system from its focus on "processing people and paper" to one that emphasizes helping recipients to become employed. From this perspective, integrating the different case management functions and assigning both to the same staff may be more effective.

To help inform this debate, the Columbus, Ohio NEWWS site conducted a social experiment that compared these two approaches in the context of a human capital oriented welfare-to work program (see Brock and Harknet, 1998). For this study, welfare staff were divided into two groups: an *integrated group* whose members performed both income maintenance and employment functions, and a *specialized group* (referred to as the "traditional" group in the original study) whose members performed either income maintenance functions or employment functions, but not both. Welfare recipients were then randomly assigned to one of these two groups or a control group that was not eligible for the program.

Interim findings from this study, based on two years of follow-up data, indicate that the integrated approach produced higher rates of program participation, although both approaches produced only modest impacts on earnings. In addition, the integrated approach produced somewhat higher welfare savings, perhaps because case managers with integrated functions discovered program noncompliance and imposed financial sanctions (which reduced welfare payments) more quickly than did those with separate functions.

- **Isolating the Effects of Financial Incentives (MFIP)**

The preceding examples illustrate how differential impact experiments can be used to determine whether one treatment strategy is better than another. This design also can be used to help untangle the effects of components of a multifaceted strategy.

One recent such application is the MDRC evaluation of the Minnesota Family Investment Program (MFIP), an innovative welfare-to-work program that combines mandated participation in employment and training services with a special financial incentive to work. The incentive was designed to "make work pay more" by allowing welfare recipients who became employed to keep more of their welfare benefits than would have been possible otherwise. An important question for policymakers in this context is whether the enhanced work incentive is sufficient by itself to markedly increase employment and earnings among welfare recipients or whether additional services are required to do so.

To answer this question, individuals were randomly assigned to either: (1) a program group that was offered only the new incentive, (2) a second program group that received the full MFIP treatment, which included both the new incentive offer and, when families had received cash welfare payments for 24 out of the past 36 months, the requirement to participate in MFIP's employment and training services, or (3) a control group, that was eligible for neither MFIP's incentive or services but could, if they met certain criteria, volunteer for alternative services (see Miller, *et al.,* 2000). By comparing subsequent outcomes for the incentives-only group with those for the control group, it was possible to estimate the *net* impacts of the financial incentive by itself. By comparing outcomes for the full-program group with those for the control group, it was possible to estimate the *net* impacts of the full program. Lastly, by comparing outcomes for the incentives-only group with those for the full-program group, it was possible to estimate the *incremental* impacts of adding MFIP services to the incentive. These findings could then be compared to corresponding estimates of program and component costs.

The results, which cover 30 months after random assignment, are most striking for long-term urban welfare recipients who were single parents. They indicate that: (1) the financial incentive offer, *by itself,* did not increase average earnings relative to the control group, although it did produce a small gain in the average rate of employment; (2) the full program increased both employment and earnings relative to the control group, and (3) both the incentive offer alone and the full program caused welfare receipt to increase relative to the control group, but the full program did so to a much lesser extent, indicating that the mandates and services it provided offset some of the effects of the financial incentive on welfare receipt (Miller, *et al.*, 2000).

- **Limitations of Differential Impact Designs**

The three-way random assignment designs used for the preceding examples represent a powerful way to compare alternative treatments or to get part way inside the black box of a multi-faceted program. However, these designs are limited by the fact that most social programs are made up of a variety of components, features, or strategies while, in most cases, it is only feasible to implement large-sample randomized social experiments with a small number of program groups. Thus, in practice, it usually is only possible to test a very small number of program components or alternative program approaches at a time. Moreover, alternative approaches often comprise multi-dimensional "packages" of features and thus also represent black boxes that, although more narrowly defined, must be unpacked themselves.

4. Pooling Experimental Data in a "Comparison-of-Sites" Analysis (Linking Program Impacts and Implementation)

As social experiments in the same field accumulate over time, new opportunities become available for combining experiments in order to test a broad range of hypotheses about what drives program performance. Specifically, *if comparable data are collected* across experimental studies and across sites within these studies, it is possible to use the natural variation in true impacts across sites (estimated by an experiment for each) and the natural variation in their programs, sample members, and environments to estimate the influence of each factor on program impacts, controlling for the others. (This comparison-of-sites design is described by Plewis (2002) in the present volume, as is the use of multi-level modeling to estimate the relationships involved. For a more extensive discussion of multi-level modeling see Bryk and Raudenbush, 1992).

The viability of this approach depends on the number of experimental sites for which comparable data have been obtained (which can be increased by pooling data across studies), the variability of true impacts across sites (the dependent variable for the analysis), the size of the experimental sample at each site (which determines the precision of site-specific program impact estimates), and the quality and consistency of measures used to characterize programs, sample members, and site environments (the independent variables for the analysis).

MDRC has just completed the first stage of a study that applies this comparison-of-experimental-sites approach to examine the effects of program management, program services, the local economy, and client characteristics on the impacts of welfare-to-work programs in a wide range of locations across the US (Bloom, *et al*., 2001). The study pools original data from three large-scale, multi-site randomized experiments conducted by MDRC: the GAIN and NEWWS evaluations discussed earlier, plus a similar evaluation of Project Independence in the State of Florida (Kemple, *et al.*, 1995). The pooled sample comprises 69,399 program and control group members from 59 local program offices (sites).

In contrast to the original evaluations, which focused on each program's overall employment, earnings and welfare impacts on *individuals*, the new study focuses on the variation in impacts across welfare-to-work *offices* that administered the programs and the factors that produced this variation. This is accomplished by estimating a two-level hierarchical linear model.

The first level of the model specifies individual outcomes (total earnings during the first two years after random assignment) as a function of individual background characteristics, program/control group status and a series of interactions between background characteristics, and program/control status. The output of this level is a program impact estimate for each local office that controls statistically for the background characteristics of its sample members. Such findings are often referred to as "conditional" impact estimates. The second level of the model specifies the conditional impact for each office as a function of how its program was managed, the extent to which it increased the use of specific employment and training services, and its prevailing local unemployment rate. Although the model is specified as two separate levels, all of its parameters are estimated simultaneously and each parameter represents the effect of a specific

individual-level or office-level variable holding constant all other variables in both levels of the model.

The office-level *program management measures* are scales (usually based on multiple items) constructed from responses to local office staff surveys conducted for the original evaluation studies. Almost all local staff in the original programs responded to these surveys and, on average, there are 21 respondents per office. From these responses, measures of the following variables were constructed: program emphasis on quick client employment; program emphasis on personalized client attention; program emphasis on close client monitoring; the differences between line staff and their immediate supervisors in the emphasis placed on this set of program practices; the variation among the line staff themselves in the emphasis placed on these practices; and average caseload size per staff member.

Office-level *program service measures* were obtained from follow-up surveys administered for the original evaluations to a subsample of program and control group members from each local office; follow-up survey samples averaged 258 respondents per office. From this information it was possible to compute separately the *difference* in the percentage of program group and control group members who received basic (remedial) education, job-search assistance, and vocational training.

Lastly, the prevailing unemployment rate for each office was obtained from publicly available sources and included in the model to represent *local economic conditions*.

Key findings from the analysis (which were subjected to an extensive series of sensitivity tests and found to be quite robust) are that: (1) A strong *employment message emphasizing quick job entry* is a powerful medium for stimulating clients to find jobs; (2) staff emphasis on *personalized client attention* markedly increases program success; (3) *large caseloads* reduce program effectiveness; (4) increased reliance on *basic education* reduces program impacts, at least in the short-term; and (5) *high unemployment rates* reduce program impacts.

These findings address major questions about welfare-to-work program design and operation and bear directly on hypotheses that have been debated for decades by practitioners and researchers alike. Hence, the findings have potentially major substantive implications for future welfare policies and programs. Furthermore, they help to illustrate what can be accomplished by pooling comparable data from a series of high-quality randomized experiments conducted in different locations and at different times. Therefore the study also has important methodological implications.

However, when assessing the comparison-of-experimental-sites approach, it is important to acknowledge that, even though evidence of each site's performance is based on a randomized experiment, all relationships between program impacts and their hypothesized determinants are estimated from a cross-sectional non-experimental analysis (the second level of the hierarchical model). Hence, these findings may be subject to the same potential for selection bias due to incomplete model specification that is inherent in any non-experimental research. The best ways to reduce the potential for such problems is to: ground the statistical model to be used as fully as possible in the theory and practice of the program studied; carefully develop and test the validity

and reliability of all measures used; and obtain a sample of sites and individuals that is as large and varied as possible. It is hoped that by pooling a series of multi-site experiments that were conducted in similar ways using comparable data that was rich in scope and based closely on widely-held (and debated) theories and hypotheses, that the present MDRC study has taken a meaningful step in this direction.

5. Using Instrumental Variables with Randomized Experiments (Exploring Causal Pathways)

One frequent complaint about randomized experiments is that they pay too little attention to the underlying theories of the programs they are designed to test. Thus, they have not contributed as much as they could have to the understanding of social problems (Chen and Rossi, 1983). To remedy this problem, what is needed, many argue, is a clear specification of the causal pathways by which impacts are expected to occur and a method of analysis that can test for the existence and strength of the intervening steps in these pathways (Weiss, 1995). Unfortunately, the effects of these intervening variables—also called mediating variables, or "mediators"—are difficult to estimate.

One promising approach for doing so is the method of instrumental variables. This approach has been well known for many years and is described in many textbooks (Greene, 1997). However, the assumptions required for the approach to provide valid results are difficult to meet in practice, and, hence, it has not played a major role in the development of program evaluation theory. Recently, however, there has been a resurgence of interest in the method, growing out of some innovative applications that use "natural experiments" to generate the exogenous variation required for the method to work (Card, 1995, and Angrist, *et al.*, 1996). Because exogenous variation also can be generated—by design—through random assignment experiments, the use of instrumental variables is now being considered within the context of such experiments as a way to study the effects of mediating variables. The basic logic of this approach is as follows. (The discussion in this section was adapted from an unpublished MDRC document prepared by Howard Bloom, with assistance from Johannes Bos, Lisa Gennetian, and Pamela Morris.)

Consider a hypothetical welfare-to-work program for adults that provides subsidized child care to parents of young children, and assume for a moment that the only way the program can affect child behavior is through its impact on the use of child care (see Figure 1a). In other words, assume that child-care is the only mediating variable in this relationship. Thus, the program impact on child behavior equals the *product* of its effect on the use of child-care times the effect of child care on child behavior. This relationship provides a simple way to use instrumental variables in the context of a random assignment experiment to estimate the effect of child-care on child behavior.

Figure 1
Illustrative Models of Mediating Variables
For Interpreting Program Impacts

Figure 1a

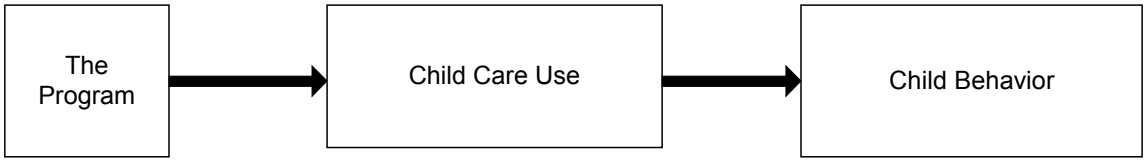


Figure 1b

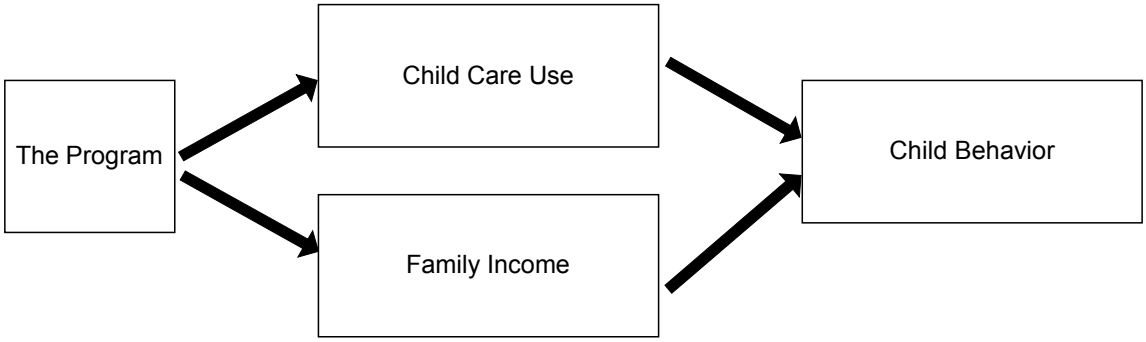
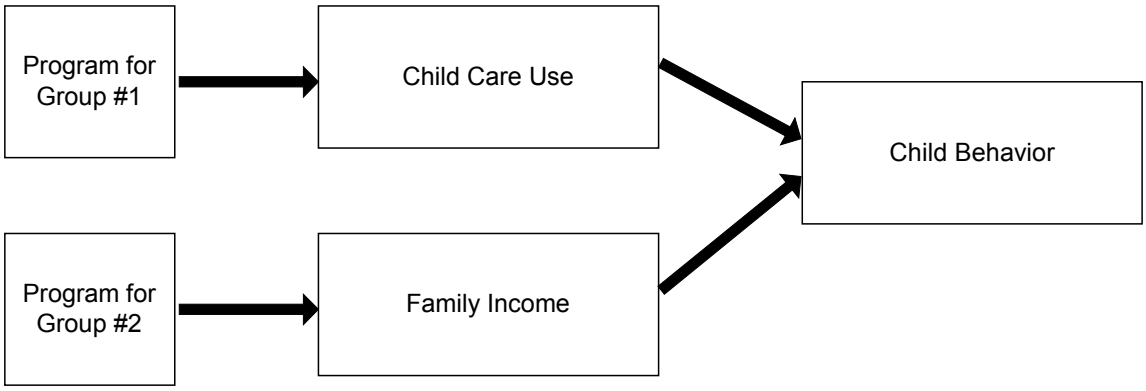


Figure 1c



To see this, note that valid estimates of two of the three elements of the relationship (program impacts on child-care use and on child behavior) can be obtained directly from an experimental design by comparing program and control group means for each outcome. Because of this, a valid estimate of the effect of child-care on child behavior can be obtained by *dividing* the estimated program impact on child behavior by the estimated program impact on child-care use. Doing so is equivalent to using the randomly assigned program status of individual sample members as an instrumental variable for estimating the relationship between child behavior and child-care; standard errors for which can be obtained readily from many software packages.

Applications of this approach to different evaluation research questions have been explored by MDRC (Bos and Granger, 1999; Morris and Gennetian, 1999; Boudett and Friedlander, 1997; and Bloom, 1984) and others (Ludwig, *et al.*, 1998). One problem that has been encountered by these applications is limited statistical power due to weak program impacts on the mediator. In other words there is often a weak connection between the mediator (use of child-care in the present example) and its "instrument" (random assignment status). This is a common problem with instrumental variables methods in both experimental and non-experimental analyses (Bound, *et al.*, 1995). To address this problem, it is necessary to identify experimental groups with large and statistically significant program impacts on the mediator of interest. MDRC is thus examining past and current experiments to identify empirical examples where this condition holds.

A second problem that has been encountered in applications of the approach is potential bias due to the existence of more than one mediator (causal pathway) for a program impact. Figure 1b illustrates this situation for a hypothetical welfare-to-work program that affects both child-care use and family income, each of which affects child behavior. In this case, there is no simple way to separate the effect of the two mediators on child behavior. If one mediator were ignored and the instrument were applied to the other, the full impact of both would be attributed to the one included. If both mediators were included in the analysis, the model would not be identified and the effects of neither mediator could be estimated.

Figure 1c illustrates an approach to dealing with the problem of multiple mediators that is being explored by MDRC—the identification of sites, subgroups or different treatment streams with a large program impact on only one mediator, with different groups having impacts on different mediators. If, for example, the program at one site had a large impact on the use of child care but not on family income (perhaps because the program focused mainly on providing child care and did little to help participants get jobs) and if there were no other plausible connection between the

program for adults and outcomes for their children, then random assignment status at this site could be used as an instrument to estimate the effect of child care on child outcomes. If, at another site, the situation were reversed—the program had a large effect on family income and not on child-care use—then random assignment status at that site could be used as an instrument to estimate the impact of family income on child outcomes. Combining the results for the two sites could thus, facilitate a comparison of the impacts of the two mediators.

At this point, several initiatives to explore these issues in the context of welfare and employment programs are underway at MDRC. In addition, other researchers are exploring applications of the approach in other settings. Although the outcomes of these efforts remain an open question, there are ample reasons for expecting them to provide new avenues for using randomized experiments to deepen our understanding of social problems and to provide insights into how programs and policies can better address these problems.

6. Combining "Cluster" Random Assignment with Interrupted Time-Series Analysis (Testing "Place-Based" Initiatives)

Recent years have seen a marked increase in comprehensive community initiatives designed to improve neighborhoods with high concentrations of poverty (Fulbright-Anderson, *et al.*, 1998). In addition, there have been several major attempts to focus public health initiatives on specific geographic areas (Murray, *et al.*, 1994). Concurrently, there have been a large and growing number of attempts to implement whole school reforms (Bloom, 2001). These initiatives share a very important common feature: they focus on whole groups, not on specific individuals.

Because of this, attempts to evaluate the initiatives share a number of common problems (Hollister and Hill, 1995; Rossi, 1999). At the root of these problems is the fact that many factors in addition to a program can spark community-wide, locality-wide, or school-wide changes. Such factors include among others: population mobility, changes in economic or social conditions, and concurrent changes in related public policies. These factors make it very difficult to isolate a specific intervention's contribution to observed change. Furthermore, because such interventions are targeted on whole groups, it is usually not feasible to randomly assign individuals to program or control status.

In some cases it is possible to use whole communities, localities, or schools (aggregate units) as the basis for analysis by randomly assigning some to the intervention and others to a control group (Boruch and Foley, 1999). This approach is referred to as "cluster" random assignment (Bloom, *et al.,* 1999, and Plewis, 2002). However, because the approach can be difficult and expensive to implement—especially for a large number of aggregate units—it is not feasible to rely *solely* on it for measuring the effectiveness of many group-focused programs. Thus, for such programs, it is important to consider alternative approaches that combine experimental and quasi-experimental methods.

MDRC is currently using such a combined strategy for a research demonstration project called The Jobs-Plus Community Revitalization Initiative for Public Housing Families (Riccio, 1999). Jobs-Plus is being sponsored by The Rockefeller Foundation, the U.S. Department of

Housing and Urban Development, and several other federal agencies and private foundations. The goal of the project, which is very ambitious, is to increase dramatically the employment rates and earnings of residents living in government-owned housing developments and thereby transform low-work, high-welfare developments into high-work, low-welfare communities.

At the heart of the Jobs-Plus approach are three core components: (1) employment-related activities and services, (2) enhanced financial incentives to work (primarily in the form of a reduction in the rent increase that normally occurs when residents increase their income by working), and (3) "community support for work", which entails, among other things, work-related information-sharing, peer support, and mutual aid among residents. Jobs-Plus is distinctive both with regard to its explicit combination of these approaches and with regard to its targeting of them on *all* working-age residents of public housing developments in the program. In other words, it attempts to "saturate" the targeted environment. It is hoped that this "saturation" approach to the provision of services, incentives, and social supports will make it possible and profitable for a substantial majority of working-age public housing residents to become employed steadily.

The project was launched in seven cities, five of which (Baltimore, Maryland; Chattanooga, Tennessee; Dayton, Ohio; Los Angeles, California; and St. Paul, Minnesota) are currently part of an impact evaluation study. Underlying the program model for this study is a theory positing that big changes in employment can only be achieved by transforming social norms and relationships within each development, which, in turn, requires reaching out to all working-age residents.

Because it was not possible to randomly assign individual residents to a program or control group, an alternative design is being used to measure the impacts of Jobs-Plus (Bloom, 1996). This design combines random assignment of selected public housing developments in each city to a program or control group (cluster random assignment) with a well-known quasi-experimental approach called "interrupted time-series analysis."

The first step in the design was for participating cities to nominate several public housing developments with roughly similar demographic characteristics (thereby matching these developments to some extent). Then for each city, one nominated development was randomly chosen to operate the Jobs-Plus program, and one or (more often) two were chosen for a control group. This random allocation of public housing developments was an attempt to avoid selection bias that would have resulted if MDRC had systematically tried to pick the "strongest" or "weakest" housing developments for the program.

Thus, when findings for all cities are pooled, subsequent differences in program and control group outcomes will provide unbiased estimates of program impacts, as is the case for any randomized experiment. However, the small number of housing developments that were randomly assigned limits the methodology in a very important way. With very few developments, it is possible that by chance, those randomly assigned to the program group will have a proportion of job-ready residents that is higher (or lower) than that of developments randomly assigned to the control group. Thus, future differences in outcomes for the two groups, while not biased *per se* (because every development in each city had the same probability of assignment to

the program) are not accurate estimates of the impacts caused by the program. If, on the other hand, many more developments had been available for random assignment, the potential for a non-comparable program and control group would be much less, and, hence, program impact estimates based on observed future outcome differences would be much more valid and reliable.

This relationship between the number of aggregate units available for random assignment and the likely validity and reliability of program impact estimates reflects the usual role of random error in statistics. For small samples the influence of such error is large and for large samples its influence is predictably small. Thus, as in any other statistical analysis, small samples have low statistical power and large samples have high statistical power. In the case of cluster random assignment, the statistical power of program impact estimates can depend far more on the number of aggregate units than on the number of individuals within each aggregate unit (Raudenbush, 1997).

Therefore, pooled experimental estimates of Jobs-Plus impacts most likely will have low statistical power, even though developments from each city were matched prior to random assignment in order to limit their initial disparities (and existing demographic and baseline survey data suggest that these disparities typically are small). This suggests that relying solely on experimental impact estimates would not be an appropriate analytic strategy for the Jobs-Plus evaluation. Further reinforcing this conclusion is the fact that experimental impact estimates for a given city are not possible because only one development per city was randomly selected to operate the program.

To address these limitations, the Jobs-Plus impact evaluation also incorporates features of a quasi-experimental design, which takes the form of a comparative interrupted time-series analysis. Interrupted time-series analysis has been used for decades to evaluate many different programs and policies (Campbell and Stanley, 1966; Cook and Campbell, 1979; Shadish, *et al.*, forthcoming, 2001). At its core, the design represents an extension of "before-after" analysis with multiple observations on outcomes before an intervention is launched (during its baseline period) and multiple observations after it is launched (during its follow-up period). In its simplest form, interrupted time-series analysis estimates program impacts for a given follow-up period as its deviation from the baseline pattern for an outcome of interest.

The primary outcomes of interest for the Jobs-Plus evaluation are employment, earnings and welfare receipt. Quarterly data on employment and earnings will be obtained from the administrative records of state Unemployment Insurance (UI) agencies for a baseline period of five years before Jobs-Plus was launched in each program development and a follow-up period of three to five years thereafter (Kornfeld and Bloom, 1999, find that UI wage records are good source of such data). Likewise, monthly data on the receipt of welfare benefit payments (for AFDC and food stamps) during the same period will be obtained from the administrative records of state or local welfare agencies. From this information, it will be possible to observe how each outcome at a given program development deviated from its baseline pattern.

To see how the analysis will work, consider the illustrative example in Figure 2. The top diagram in the figure illustrates how, during each follow-up quarter, employment rates at a hypothetical Jobs-Plus development deviated from its baseline trend line. This is the first part of

the estimation strategy. The example in the figure indicates that employment rates increased markedly beyond that predicted by the baseline trend—which is what is hoped for.

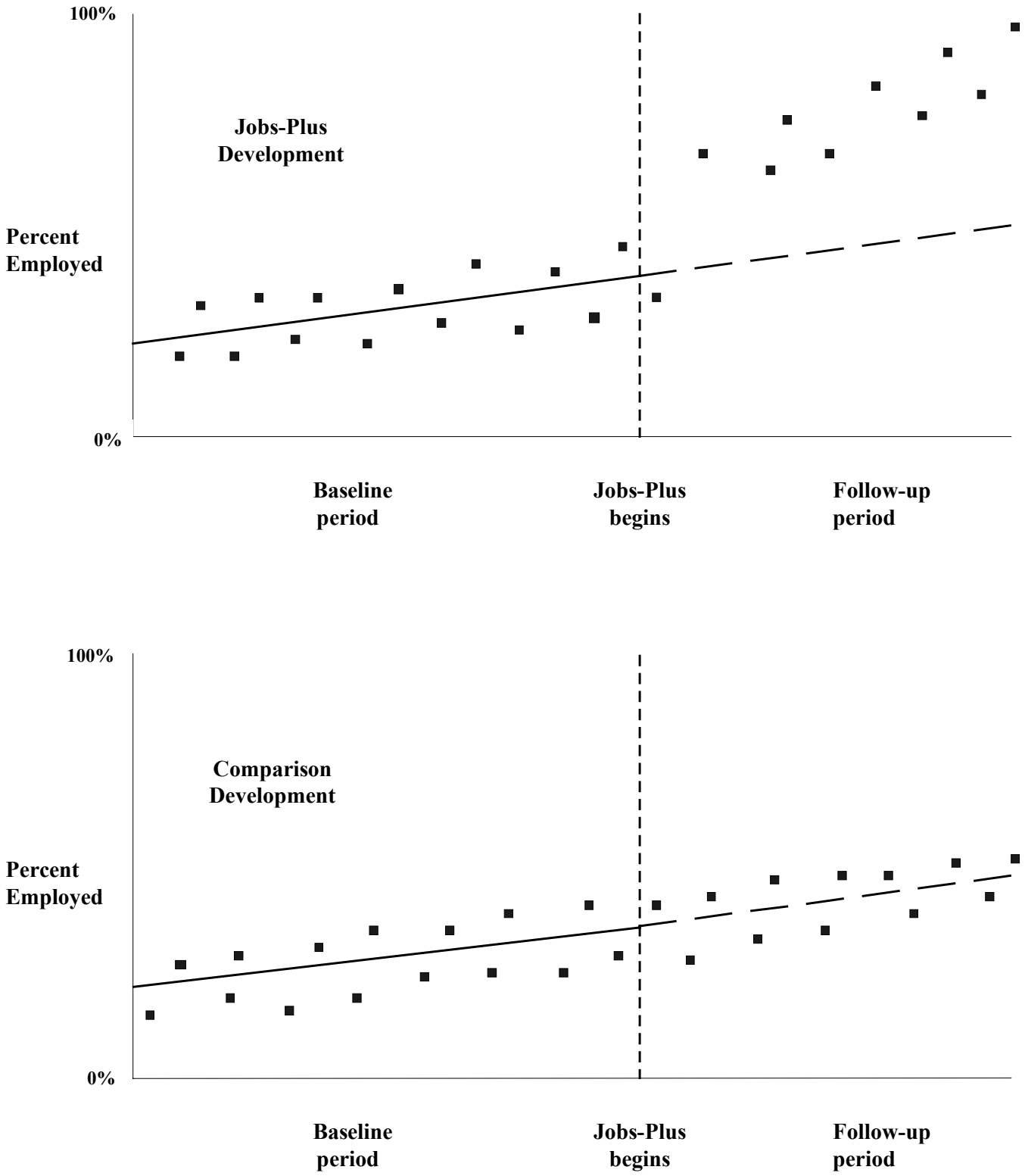
The next part of the estimation strategy involves a corresponding analysis for a comparison/control development. Results of this analysis can be used to predict what the deviation from trend would have been for the Jobs-Plus development if the program had not been launched there. The bottom diagram in Figure 2 illustrates that there was little or no deviation from the baseline trend at the hypothetical comparison development—which is what might be expected under normal economic conditions.

The final part of the estimation strategy involves taking the difference between the deviation from trend at the Jobs-Plus development and the deviation from trend at the comparison development for each segment of the follow-up period (calendar quarters in the present example). Thus program impacts are estimated as "differences of deviations from trend," which is similar to in spirit but different in form from a standard difference-of-differences estimator.

The preceding approach makes it possible to estimate the impacts of Jobs-Plus for each site. In addition, when impact estimates are pooled across sites, they have the combined methodological protection of interrupted time-series analysis *and* cluster random assignment. Therefore this approach may represent a promising candidate for assessing the effectiveness of other community initiatives.

Figure 2

A Comparative Interrupted Time-Series Analysis of Jobs-Plus Impacts on Employment Rates



Nevertheless, in principle, the approach is weaker than a classical randomized experiment. Thus, for program evaluations that must identify small impacts with confidence, the approach may not be adequate. However, the explicit goal of Jobs-Plus is to produce *large* impacts. From a policy perspective, it will not have succeeded if it only produces small effects. Therefore knowing whether small impacts are "real" is not essential.

7. Conclusions

This paper has described the evolution of social experiments used to evaluate welfare-to-work and employment programs. These experiments have developed from simple "black box" designs that only measure the "net" impact of an overall program to more complex multi-group designs that measure the "differential" impacts of alternative program strategies or the "incremental" impacts of adding program elements to each other. In addition, evaluators have begun to augment social experiments with non-experimental and quasi-experimental strategies. Many of the studies reviewed above are still underway or just beginning, so the full contribution of the methods they employ remains to be seen. However, we believe that they reflect an important push in new directions that ultimately will extend the reach of social experiments and thus increase their contribution to knowledge.

Thus, future evaluation research could do well to continue down this path. In some cases, this means designing more sophisticated social experiments from the start. In other cases it means envisioning the potential scientific value of pooling data from different social experiments and ensuring that the common types of information necessary to do so are collected. In yet other cases, it means looking within experiments for opportunities presented by the pattern of findings that can be exploited to test important hypotheses (as in the use of instrumental variables). Furthermore, it also means thinking creatively about ways to combine experiments with other methods in order to address important evaluation questions they cannot be studied by classical experimental methods alone.

In closing, it is important to note that the debate over the advantages and disadvantages of experimental and non-experimental methods is longstanding, vigorous and remains important. However, dwelling only on the likely superiority of one approach over the other is likely to be counterproductive. Thus, instead of seeking to inspire wholesale acceptance or rejection of one set of methods over the other, evaluation researchers should engage in a more intensive search for new ways and opportunities to use both productively.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*. Vol. 91, pp. 444-472.

Betsey, C., Hollister, R. and Papageorgiou, M. (1985) *Youth Employment and Training Programs: The YEDPA Years.* Committee on Youth Employment Programs, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, D.C.: National Academy Press.

Bloom, H. S. (1984) "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 8 (April): 225-46.

Bloom, H. S. (1996) *Building a Convincing Test of a Public Housing Employment Program: Planning for the Jobs-Plus Demonstration.* (Reissued in 1999 as a working paper.) New York: Manpower Demonstration Research Corporation.

Bloom, H. S. (2001) *Measuring the Impacts of Whole School Reforms: Methodological Lessons from an Evaluation of Accelerated Schools.* New York: Manpower Demonstration Research Corporation, March.

Bloom, H. S., et al. (1993) *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months.* Bethesda: Abt Associates, Inc.

Bloom, H. S., Bos, J. M. and Lee, S. W. (1999) "Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs." *Evaluation Review*, Vol. 23, No. 4, pp. 445-469.

Bloom, H. S., Hill, C. J. and Riccio, J. A. (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.

Boruch, R. F. (1997) *Randomized Experiments for Planning and Evaluation: A Practical Guide.* Vol. 44. Thousand Oaks, CA: Sage Publications.

Boruch, R. F. and Foley, E. (1999) "The Honestly Experimental Society: Sites and Other Entities as the Units of Allocation and Analysis in Randomized Experiments." Philadelphia: Center for Research and Evaluation in Social Policy, University of Pennsylvania.

Bos, J. M. and Granger, R. C. (1999) *Estimating Effects of Day Care Use on Child Outcomes: Evidence from the New Chance Demonstration.* New York: Manpower Demonstration Research Corporation.

Boudett, K. P. and Friedlander, D. (1997) "Does Mandatory Basic Education Improve Achievement Test Scores of AFDC Recipients: A Reanalysis of Data from California's GAIN Program." *Evaluation Review*, 21 (5): 568-588.

Bound J., Jaeger, D. and Baker, E. (1995) "Problems with Instrumental Variables

Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association*, Vol. 90, pp. 443-450.

Brock, T. and Harknett, K. (1998) "A Comparison of Two Welfare-to-Work Case Management Models." *Social Service Review,* 74 (2): 494-520.

Bryk, A. and Raudenbush, S. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.

Campbell, D. T. and Stanley, J. C. (1966) *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally).

Card, D. (1995) "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*. (eds L.N. Christofides, E. I. Grant, and R. Swidinsky), Toronto: University of Toronto Press.

Chen, H. and Rossi, P. H. (1983) "Evaluating with Sense: the Theory-Driven Approach," *Evaluation Review*, Vol. 7, pp. 283-302.

Cook, T. D. and Campbell, D. T. (1979) *Quasi-Experimental Design and Analysis Issues for Field Settings* (Chicago: Rand McNally).

Freedman, S., Friedlander, D., Lin, W. and Schweder, A. (1996) "The GAIN Evaluation: Five-Year Impacts on Employment, Earnings, and AFDC Receipt." Paper. New York: Manpower Demonstration Research Corporation.

Freedman, S., Knab, J., Gennetian, L. A., and Navarro, D. (2000) *The Los Angeles Jobs-First GAIN Evaluation: Final Report on a Work First Program in a Major Urban Center.* New York: Manpower Demonstration Research Corporation.

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

Friedlander, D., Greenberg, D. and Robins, P. K. (1997) "Evaluating Government Training Programs for the Economically Disadvantaged." *Journal of Economic Literature*. 35: 1809-55.

Fulbright-Anderson, K. Kubisch, A. C. and Connell, J. P. (1998) *New Approaches to Evaluating Community Initiatives: Volume 2, Theory, Measurement and Analysis* (Washington, DC: The Aspen Institute).

Greenberg, D. and Shroder, M. (1997) *Digest of Social Experiments*, 2d ed. Washington, D.C.: Urban Institute Press.

Greene, W. H. (1997) "Instrumental Variables Estimation," *Econometric Analysis*, 3rd Edition. Upper Saddle River, NJ: Prentice Hall. pp. 288-297.

Gueron, J. M. (1999*) The Politics of Random Assignment: Implementing Studies and Impacting Policy.* New York: Manpower Demonstration Research Corporation.

Gueron, J. M. and Pauly, E. (1991) *From Welfare to Work*. New York: Russell Sage Foundation.

Hamilton, G., Brock, T., Farrell, M., Friedlander, D. and Harknett, K. (1997) *National Evaluation of Welfare to Work Strategies: Evaluating Two Welfare-to-Work Program Approaches: Two-Year Findings on the Labor Force Attachment and Human Capital Development Programs in Three Sites.* Washington, D.C.: U.S. Department of Health and Human Services, Administration for Children and Families and Office of the Assistant Secretary for Planning and Evaluation.; and U.S. Department of Education, Office of the Under Secretary and Office of Vocational and Adult Education.

Hollister, R. G., and Hill, J. (1995) "Problems in the Evaluation of Community-Wide Initiatives." In *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts.* (eds James P. Connell et al.), Washington, D.C.: Aspen Institute.

Kemple, J., Friedlander, D. and Fellerath, V. (1995) *Florida's Project Independence: Benefits, Costs, and Two-Year Impacts of Florida's JOBS Program*. New York: Manpower Demonstration Research Corporation.

Kornfeld, R. and Bloom, H. S. (1999) "Measuring Program Impacts on Earnings and Employment: Do Unemployment Insurance Wage Reports from Employers Agree with Surveys of Individuals?" *Journal of Labor Economics*. 17 (1), 168-197.

Miller, C., Knox, V., Gennetian, L. A., DoDoo, M., Hunter, J. and Redcross, C. (2000) *Reforming Welfare and Rewarding Work: Final Report on the Minnesota Family Investment Program, Volume 1, Effects on Adults*. New York: Manpower Demonstration Research Corporation.

Morris, P. A. and Gennetian, L. A. (1999) "Identifying Causal Effects of Poverty on Children's Development: Integrating an Instrumental Variables Analytic Method with an Experimental Design." Paper. New York: Manpower Demonstration Research Corporation.

Murray, D. M., Hannan, P. J., Jacobs, D. R., McGovern, P. J., Schmid, L. , Baker, W. L. and Gray, C. (1994) "Assessing Intervention Effects in the Minnesota Heart Health Program, *American Journal of Epidemiology* 139 (1): 91- 103.

Plewis, I. (2002) "Modeling Impact Heterogeneity" *Statistics in Society***.**

Raudenbush, S. W. (1997) "Statistical Analysis and Optimal Design in Cluster Randomized Trials." *Psychological Methods* 2 (2) 173-85.

Riccio, J. A. 1999. *Mobilizing Public Housing Communities for Work: Origins and Early Accomplishments of the Jobs-Plus Demonstration.* New York: Manpower Demonstration Research Corporation.

Riccio, J. A., Bloom, H. S. and Hill, C. J. (2000) "Management, Organizational Characteristics, and Performance: The Case of Welfare-to-Work Programs." In *Governance and Performance: Models, Methods, and Results* (eds L. E. Lynn, Jr. and C. Heinrich), Georgetown University Press.

Riccio, J. A., Friedlander, D. and Freedman S. (1994) *GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program.* New York: Manpower Demonstration Research Corporation.

Riccio, J. A., and Hasenfeld, Y. (1996) "Enforcing a Participation Mandate in a Welfare-to-Work Program." *Social Service Review*, 70 (4): 516-42.

Riccio, J. A., and Orenstein, A. (1996) "Understanding Best Practices for Operating Welfare-to-Work Programs." *Evaluation Review*, 20 (1): 3-28.

Rossi, P. H. (1999) "Evaluating Community Development Programs: Problems and Prospects." In *Urban Problems Community Development*. (eds R. F. Ferguson and W. T. Dickens), Washington, D.C.: Brookings Institution Press.

Shadish, W. R., Cook, T. D and Campbell, D. T. (forthcoming, 2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Weiss, C. H. (1995) "Nothing As Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families." In J. Connell, A. C. Kubisch, L. B. Schorr, and C. H. Weiss, *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts* Washington, DC: The Aspen Institute.

About MDRC

The Manpower Demonstration Research Corporation (MDRC) is a nonprofit, nonpartisan social policy research organization. We are dedicated to learning what works to improve the well-being of low-income people. Through our research and the active communication of our findings, we seek to enhance the effectiveness of social policies and programs. MDRC was founded in 1974 and is located in New York City and San Francisco.

MDRC's current projects focus on welfare and economic security, education, and employment and community initiatives. Complementing our evaluations of a wide range of welfare reforms are new studies of supports for the working poor and emerging analyses of how programs affect children's development and their families' well-being. In the field of education, we are testing reforms aimed at improving the performance of public schools, especially in urban areas. Finally, our community projects are using innovative approaches to increase employment in low-income neighborhoods.

Our projects are a mix of demonstrations — field tests of promising program models — and evaluations of government and community initiatives, and we employ a wide range of methods to determine a program's effects, including large-scale studies, surveys, case studies, and ethnographies of individuals and families. We share the findings and lessons from our work — including best practices for program operators — with a broad audience within the policy and practitioner community, as well as the general public and the media.

Over the past quarter century, MDRC has worked in almost every state, all of the nation's largest cities, and Canada. We conduct our projects in partnership with state and local governments, the federal government, public school systems, community organizations, and numerous private philanthropies.