MDRC Working Papers on Research Methodology

Using Instrumental Variables Analysis to Learn More from Social Policy Experiments

Lisa A. Gennetian
Johannes M. Bos
Pamela A. Morris

MDRC

Manpower Demonstration
Research Corporation

October 2002

This working paper is part of a series by MDRC on alternative methods of evaluating the implementation and impacts of social programs and policies.

Funding for the paper was provided by the Pew Charitable Trusts through a grant to support MDRC's Methodological Innovations Initiative and by the Russell Sage Foundation through a grant to support MDRC's preparation of a book exploring how to combine experimental and nonexperimental methods for policy research.

The authors thank the following people for their valuable feedback and input on the paper: David Card (University of California at Berkeley), Greg Duncan (Northwestern University), Guido Imbens (University of California at Berkeley), Bruce Meyer (Northwestern University), Howard Bloom (MDRC), and Charles Michalopoulos (MDRC).

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 Atlantic Philanthropies; the Alcoa, Ambrose Monell, Bristol-Myers Squibb, Fannie Mae, Ford, George Gund, Grable, New York Times Company, Starr, and Surdna Foundations; and the Open Society Institute.

The findings and conclusions presented are those of the authors and do not necessarily represent the positions of the project funders or advisors.

For information about MDRC, see our Web site: www.mdrc.org.
MDRC® is a registered trademark of the Manpower Demonstration Research Corporation.

Abstract

One strategy for discovering the connections between social policy interventions and behavioral outcomes is to conduct social experiments that use random assignment research designs. Although random assignment experiments provide reliable estimates of the effects of a particular policy, they do not reveal how a policy brings about its effects. If policymakers had answers to the "how" questions, they could design more effective interventions and make more informed policy trade-offs. This paper reviews one promising approach to specifying the causal paths by which impacts are expected to occur: instrumental variables analysis, a method of estimating the effects of intervening variables — also called mediating variables, or mediators — that link interventions and outcomes. It explores the feasibility of applying this approach to data from random assignment designs, reviews the policy questions that can be answered using the approach, and outlines the conditions that have to be met for the effects of mediating variables to be estimated. Illustrations of instrumental variables analysis based on data from random assignment studies are also presented.

Contents

List of Tables, Figures, and Boxes

Table

Figure

Box

Introduction

Because so many factors influence human behavior, making a clear link between behavioral influences and behavioral outcomes is anything but straightforward. One strategy for making such links is to conduct social experiments that use random assignment research designs to answer questions about the effects of social policy interventions. The use of random assignment in such experiments eliminates most common sources of bias from these estimates, producing findings that are largely undisputed and easy to interpret (see Robins and Greenberg, 1986; Orr, 1999, for a review). By assigning individuals at random to treatment and control groups, any difference between the two groups can be attributed to the treatment.

In principle, random assignment experiments can be designed to answer any social policy question. In practice, however, random assignment experiments have important limitations. First, these experiments require a well-controlled and well-defined "counterfactual" state. (The counterfactual is the condition that would have existed in the absence of the policy intervention or program.) This counterfactual state, to which control group members in the experiment are assigned, determines and limits which policy questions the random assignment experiment can answer, as the effect of the intervention is always determined *relative* to this counterfactual. Second, the policy question being studied has to be "assignable," meaning that enrollment in a program or exposure to a "treatment" is manipulated through an external mechanism that is part of the research design. Often, policy questions cannot be manipulated that way. Several researchers, for example, have attempted to measure the effects of receiving a General Educational Development certificate (GED) on those who earn this alternative high school credential (Bos et al., 2001; Cameron and Heckman, 1998; Tyler, Murnane and Willet, 2001), but since a GED is earned, and cannot be assigned by researchers, all of these studies relied on nonexperimental comparisons of GED-holders and others. Third, nonparticipation and varying levels of participation in a treatment can affect the estimates of the effect of the treatment. A recent study, for example, examined the relationship between the number of months that welfare recipients spent in adult education programs and their subsequent employment outcomes (Bos et al., 2001). While welfare recipients in this study were randomly assigned to an "adult education" program stream, many did not participate, and those who did participate received education services for widely varying amounts of time. Both of these decisions were beyond the control of the researchers despite the underlying experimental research design. As a result, the estimates of the effects of the program were attenuated due to nonparticipation, and estimates of the relationship between the amount of participation and subsequent outcomes were strictly nonexperimental.

Finally, experimental designs do a very good job of providing estimates of the effects of a particular policy, but not about how those effects occurred. Yet many researchers are interested in *how* programs that are evaluated using random assignment designs achieve their effects

on those assigned to them. Answering these "how" questions can help policymakers design more effective interventions and can help them make difficult policy trade-offs. For example, recent evaluations of welfare and work programs have found that some of these programs may improve school outcomes for children in elementary school (Morris et al., 2001). However, an open question is whether these effects are driven by the increased financial and other resources available to program participants (since increases in family income are believed to benefit children) or by the increased employment among their parents (since working parents function as role models for their children, possibly increasing children's motivation to do well in school, and employment may also enhance the regularity of family routines). Knowing how these components work together to produce the desired effects could inform funding decisions that trade off resources spent on services such as child care and case management and resources channeled directly to low-income families. These "how" questions about the overall program effects are often considered part of a "black box" of program effects, a term that reflects the common perception that these questions are essentially unanswerable or at least very difficult to address.

This chapter reviews one promising approach for clearly specifying the causal paths by which impacts are expected to occur; a method of analysis that can estimate the effects of intervening variables — also called mediating variables, or "mediators." Subsequent sections explore the feasibility of this approach using data from random assignment designs, review the policy questions that can be answered using such an approach, and introduce the conditions that have to be met to estimate the effects of mediating variables.

Instrumental Variables Analysis as a Nonexperimental Alternative to Understanding Program Impacts

Nonexperimental research methods can be used to address policy questions that are not readily tested using random assignment or are concerned with how programs produce their effects. Such methods include *cross-sectional* comparisons of outcomes across different levels of a policy variable in a sample, *longitudinal* analyses of changes in those outcomes over time, or combinations of the two. When examining the causal effect of one variable, researchers face the need to isolate the particular variable of interest from other important variables that are correlated with it. Thus, going back to an earlier example, in comparing the outcomes of GED holders with those of GED nonholders, they need to account for social background variables that may cause some people to pursue such a credential while others do not. In the following model,

Y_i is the outcome, the background variables are shown as Z_{ik}, and the policy variable of interest is labeled X_i.[1]

$$Y_i = \alpha + \beta_i X_i + \sum_{1}^{k} \beta_{ik} Z_{ik} + \varepsilon_i \quad (1)$$

Including the background variables Z_{ik} is intended to account for the fact that GED-holders and those without a GED have different background characteristics, and controlling for them should improve estimates of the causal effect of having a GED. However, though theory can help determine which background variables should be included in an empirical estimation, some of these background variables are unobservable or difficult to measure, such as motivation. Consequently, it is virtually impossible to have all the information necessary — that is, all the "theoretically-motivated" background variables — available to include in the empirical estimation. This is a challenge that faces anyone who conducts nonexperimental research. It is always possible that a key explanatory variable Z_i is left out of the analysis. If that is the case, its effect on the outcome may be misattributed to the policy variable X_i.

Many books and articles have been written about the limitations of nonexperimental research methods, most of which have to do with researchers' inability to take account of all alternative explanations for the empirical relationships they observe in nonexperimental data. These limitations affect both types of nonexperimental research being conducted. In cross-sectional research, failure to account for important alternative explanations is commonly known as "selection bias," and in longitudinal research such a failure creates "history" or "maturation" bias (see Cook and Campbell, 1979, for an extensive discussion of these limitations of nonexperimental research).

Two distinctive techniques — "differencing" with data from natural experiments and *instrumental variables analysis* — represent common approaches to capitalizing on exogenous variation in policy variables to control for these biases in understanding empirical relationships.

Natural Experiments

So-called "natural experiments" are the most common, and arguably the most intuitive, technique. In a natural experiment, researchers take advantage of a situation in which two otherwise identical groups (or time periods) are affected differently by a "natural" event P_i that is exogenous to the relationship between Y_i and X_i and causes a sufficiently large change in both.

[1] X_i is called a policy variable throughout this paper because it captures a well-understood construct that policymakers might want to manipulate, such as educational attainment, child care use, or family income. However, in almost every case discussed here, it is not feasible to apply random assignment to X_i itself.

In a famous example in the minimum wage literature, Card and Krueger (1994) compared the labor demand of fast food restaurants in the border region of Pennsylvania and New Jersey after the minimum wage was raised in New Jersey. Contrary to the prevailing theory, they found no evidence that this policy decision reduced the demand for fast food labor. Their results were convincing because economic and demographic conditions within their cross-state sample were virtually identical *except for the change in the minimum wage law.*[2] The presence of a state border and the difference in state policy created a natural experiment to test the effect of wage policy on the demand for low-wage labor.

There are numerous examples of research that uses such "natural" variation in policy variables to produce unbiased estimates of the effects of those policy variables on human behavior and economic conditions. For example, Hotz et al. (1999) estimated the effects of teen childbearing by comparing outcomes for teens who had miscarriages with outcomes for teens who carried their children to term. Hoxby (2001) estimated the effects of vouchers on school choice using school district boundaries determined by streams. And, recently, Angrist (forthcoming, 2002) estimated the effects of sex ratios on marriage and labor market outcomes using immigration inflows into the United States.

Instrumental Variables Estimation

Another powerful nonexperimental technique for addressing questions of causality is the use of instrumental variables (IV) estimation strategies. These strategies rely on finding an independent ("exogenous") source of random variation in the policy variable X_i whose effects are being analyzed. This exogenous variable (hereafter referred to as P_i) is known as the "instrument." In a simple instrumental variables framework, the effect of X_i on outcome Y_i is estimated by comparing the effect of P_i on Y_i to the effect of P_i on X_i, or more explicitly:

$$\frac{dY_i}{dX_i} = \frac{\dfrac{dY_i}{dP_i}}{\dfrac{dX_i}{dP_i}}, \tag{2}$$

where $\frac{dY_i}{dX_i}$ is the effect of X_i on Y_i, $\frac{dY_i}{dP_i}$ is the effect of P_i on Y_i, and $\frac{dX_i}{dP_i}$ is the effect of P_i on X_i.

[2]While there was little criticism of the underlying research design, some researchers did question other aspects of this study, such as data collection and sample selection (Wascher and Neumark, 1992, 1994).

Limitations of Nonrandomized Instruments

These two techniques illustrate that to produce valid findings, the variation in the policy variable X_i does not have to be random as in a real experiment provided that this variation is exogenous to outcome Y_i. More formally, this is expressed as follows:

$$Y_i = \alpha + \beta_i X_i + \sum_{1}^{k} \beta_{ik} Z_{ik} + \varepsilon_i \text{ , where} \tag{3}$$

$$X_i = \gamma + \delta_i P_i, \tag{4}$$

$$COV(P_i, Z_i) = 0 \text{ , and} \tag{5}$$

$$COV\left|(P_i, {}_i X_i)\right| > 0 \text{ , and} \tag{6}$$

$$COV(Y_i, P_i \mid X_i) = 0. \tag{7}$$

Remember, in such a system of equations, P_i is the "instrument" for X_i. The equation in which the instrument is used to predict X_i is called the first-stage equation (with the second-stage equation being that in which the predicted value of X_i is used to predict Y_i). Because all possible variables Z_i are not known, one cannot be certain that COV(P_i,Z_i) is indeed zero, unless P_i is a random variable.

This observation highlights one of the limitations of natural experiments. Although natural experiments add valuable independent variation into analyses of effects of policy variables on outcomes, it is almost never possible to guarantee that P_i is uncorrelated with all unmeasured variables Z_i. For example, in cases where "natural" policy variation across jurisdictional boundaries is studied, researchers must assume that people are randomly distributed on different sides of the boundary or at least that their choice to live on one side or the other is uncorrelated with the policy variable X_i. Given that many of these policy variables are affected by systematic preferences or differential ability to choose where to live (e.g., some people may not have the economic means to live on one side of the boundary), it is often difficult to make a convincing argument that P_i is indeed uncorrelated with Z_i. Moreover, it is often very difficult to find natural experiments to answer critical policy questions. Using natural experiments requires a great deal of opportunism on the part of researchers, which often leads to compromises in other aspects of the research, such as the study's generalizability, the quality of the available data, or the researchers' ability to measure P_i with precision. The most highly regarded examples of "natural" experiments in policy research are those that use true lotteries that are implemented for nonresearch reasons. In the labor economic literature, many studies rely, for example, on the Vietnam era draft lottery, which constituted a strong exogenous incentive to change

one's behavior for many of those affected (e.g., enroll in college, have a family, volunteer for the army, or move to Canada). Studies based on this lottery were used to estimate the economic benefits of postsecondary education (Angrist and Krueger, 1992). However, drastic lottery-based "treatments" like the Vietnam era draft lottery are very rare, which means that researchers who rely on natural experiments rarely have such convincing instruments to work with.

Random Assignment as an Instrument

Since a variable is only a "good" instrument if it is *known* a priori to be uncorrelated with any unmeasured explanatory variables Z_{ik}, the best instruments are those whose values are assigned randomly. Randomized experiments are designed to create a "program" variable that has randomly assigned values. Provided that a program variable has a meaningful relationship with a policy variable of interest, it is a natural choice of instrument for this policy variable. Thus, for example, if researchers studying the effects of the GED on earnings could identify an experimental treatment that affected GED receipt (such as a program that promoted taking the GED test among high school dropouts with sufficient skills to pass it), they could, in theory, use the experimental treatment variable as an instrument for GED receipt. In the remainder of this paper, we will explore the assumptions under which such an approach would be valid, its limitations in terms of statistical power and data requirements, and the potential to develop studies that would use random assignment expressly as a way to generate valid and powerful instruments to address important policy questions.

Policy Questions Answered by IV and the Assumptions Needed to Answer These Questions

What Policy Questions Can IV Answer When Combined with Experimental Designs?

Using IV with random assignment allows researchers to answer a wider range of policy questions than are answered by random assignment studies alone. This section describes the range of questions that can be addressed when random assignment studies are combined with IV estimation strategies. Moreover, this section demonstrates how instrumental variables estimators can help make the answers from random assignment studies speak more directly to the policy questions that researchers are ultimately trying to address (see Angrist, Imbens, and Rubin, 1996, for further discussion). For definitions of the key terms used in this section, see the text box on page 8.

An "Intention" to Treat

When a random assignment experiment is conducted, the difference in outcomes between those assigned to the program group and those assigned to the control group is a fully experimental (and valid) estimate of the program effect. However, in many cases, the "program" whose effect is captured is not equivalent to the policy being studied. For example, suppose a training program aims to provide 30 weeks of vocational skills training to those who enroll. A sample of applicants for the program is recruited, random assignment is conducted, and half of the sample is offered the program. Many of those randomly assigned to the program group follow through and indeed participate for 30 weeks. However, some drop out, and others do not show up at all. As a result, the program-control group difference at the end of the study does *not* constitute the effect of 30 weeks of vocational skills training. Instead, it captures the effect of a program's *intention* to provide such a level of training. Among economists, this effect is known as the "intent-to-treat" (or ITT) effect. The relevance of the ITT estimate for policymakers depends on the rate of take-up in the program and the initial policy question. Critics of random assignment studies often point to the limitations of ITT estimates as a major drawback of answering policy questions with random assignment research (e.g., see Heckman, 1997).

The Effect of Receiving the Treatment

Many of these critics argue that a more relevant measure is the effect of "treatment on the treated" (TOT). In the example above, this measure would capture the effect of actually receiving the 30 scheduled weeks of training (the "treatment" in the context of that example). It answers the question of how those who received the training benefited from their experience. However, there are two serious problems with the TOT effect. First, it is very difficult to estimate, because it is difficult to establish a priori who is going to be among the "treated." (If that were possible, the experiment could just be limited to that group.) Second, even if it were possible to reliably estimate a TOT effect, it would be difficult to generalize it to a wider policy. Knowing, for example, that 30 weeks of vocational skills training produces an X percent increase in the earnings of those who received it does not answer the question of how to go about getting those who need higher earnings to participate in such training.

Instrumental variables estimation strategies can be used to approximate TOT effects. The coefficient β_i in Equation 3 is often considered an instrumental variables estimator of the TOT effect of X_i on Y_i. In that example, the instrument P_i is the program (or the "intention to treat") and X_i is the "treatment." However, as Angrist et al. (1996) point out, the effect β_i is not a true TOT effect but a special case of such an effect, which they define as the "local average treatment effect" (LATE). Whereas TOT is the general effect of a treatment on those who receive it, LATE is defined more precisely as the effect of a "treatment" on those who are *induced*

by a specific program to receive it.[3] In experimental research, these individuals are also known as "compliers" (sample members who would not have received the "treatment" had they been assigned to the control group; Angrist et al., 1996). A key difference between LATE and TOT is that the latter applies to a clearly defined and identifiable subpopulation, namely, those who receive the treatment (Heckman, 1996a, 1996b). (In the example above, these are all the sample members for whom $X_i = 1$, regardless of the value of P_i.) The LATE estimator is limited to compliers, who cannot be identified either *a priori* or *ex post facto* in any experiment.

Definition of Terms

ITT (Intent to Treat). ITT captures the effect of a program's intention to provide a certain level of services or benefits.

TOT (Treatment on the Treated). TOT captures the effect of a program on those who received it.

LATE (Local Average Treatment Effect). LATE captures the effects of a program on those who were induced to receive it, that is, compliers.

ATE (Average Treatment Effect). ATE generalizes LATE a broader population of compliers.

Uses of the Local Average Treatment Effect

The relevance of LATE estimators is hotly debated. In support of LATE estimators, one might argue that it is especially helpful to understand the effects of a policy variable on those for whom this variable can be manipulated through an external program (as opposed to those for whom such manipulation makes no difference). However, the usefulness of LATE estimators is limited by researchers' inability to understand or control the process of compliance. As Robins and Greenland (1996) argue, if an experiment shows that a treatment X_i is beneficial, making X_i widely available in the population will result in a different take-up pattern than that found in the experiment that produced the LATE estimator of X_i's effect. For example, many more people may take aspirin after a study's results show it to be effective in treating heart attacks. Therefore, there will probably be more — and different — compliers. The actual effect that would take this into account, which Robins and Greenland refer to as the "average treatment effect"

[3]Note that the term treatment is easily confused with the experimental status. In discussions of TOT and LATE effects, the "treatment" is the policy variable referred to earlier as X_i, not the experimentally manipulated program variable P_i.

(ATE), cannot be estimated directly, but they demonstrate that it is possible to develop bounds for this effect, given valid estimates of the ITT and LATE effects.[4]

Given a number of strict assumptions, which will be discussed below, Angrist et al. (1996) show that the instrumental variables estimator β_i is a valid LATE estimator of the effect of X_i on Y_i. It is not a valid TOT estimator of X_i on Y_i, unless everyone who receives the treatment X_i does so by way of program P_i. But, as discussed above, LATE estimators are policy relevant because they capture the effects of externally induced changes in policy variables, which is the purpose of much social policy research. The following section describes these assumptions and discusses both the likelihood that they are violated in real-life situations and the consequences such violations would have for the validity of estimated effects.

Assumptions to Identify Causal Effects

Consider again the instrumental variables estimator presented in Equations 2 and 3. There is a randomized program variable P_i, which produces a change of δ_i in the policy variable X_i. The change in outcome Y_i associated with this randomly induced change in X_i is captured by the instrumental variables estimator β_i , which is a valid LATE estimator of the effect of X_i on Y_i if the following five assumptions are met (Angrist et al., 1996; for a list, see the text box on page 11).

First, the instrument P_i is assumed to be a randomly assigned variable, meaning that it is uncorrelated with demographic characteristics of persons i, with preprogram levels of outcome Y_i, and with any other preprogram variables Z_i that could predict Y_i. As mentioned above, instead of being randomly assigned, it would be sufficient for P_i to be a nonrandom but truly exogenous variable, provided that it were possible to demonstrate that it was.

Second, there must be a meaningful effect of the instrument P_i on the policy variable X_i, i.e., $\delta_i \neq 0$. In later sections of this chapter, it will become clear that the larger the program effect δ_i is, the more reliable the instrumental variables estimator β_i will be. When δ_i is small, P_i is said to be a "weak" instrument.

The third assumption is referred to by Angrist et al. (1996) as the "stable unit treatment value assumption" (SUTVA). This assumption requires that the values of policy variable X and the relationships between those policy values and outcome values Y across individuals i are "stable" for those individuals, that is, unaffected by variation in X or Y for other individuals. Without this assumption, it is impossible to draw reliable inferences about the effects of the policy variable on the outcome, regardless of the type of statistical analysis that is used. In practice,

[4]Robins and Greenland (1996) present an extensive review of the biomedical literature to support these bounds.

this means that there are assumed to be no community effects or displacement effects; *Y, X*, and *P* are independent across different individuals *i.*

A violation of SUTVA creates a bias in the estimator β_i . The size and direction of this bias depend on the nature and seriousness of the violation. For example, consider the hypothetical training program introduced earlier. Say there is a limited number of skilled positions in a geographic area and program P dramatically increases the number of persons holding a training credential. As a result, employers are less willing to pay high wages to retain credentialed persons, that is, the premium associated with a credential is reduced. In this case, the increased level of educational attainment (X_i) in the community, associated with program P, has lowered earnings for those who would have held a credential even without program P. As a result, the estimated effect β_i is an underestimate of the true benefit associated with receipt of a credential. Fortunately, it is reasonable to assume that SUTVA holds in most cases as the size of programs tends to be small relative to the communities in which they are implemented.

A fourth assumption is that the program effect δ_i on the policy variable (i.e., the effect of P_i on X_i) is monotonic. This means that when $\delta_i > 0$, ($X_i|P_i=1$) is greater than or equal to ($X_i|P_i=0$) for every person *i.* Thus, for example, if P_i is a training program and X_i is the amount of training received, no one randomly assigned to the program receives less training than s/he would have received if s/he had *not* been assigned to the program. The consequences of violating this assumption depend on the relationship between X_i and Y_i for compliers and "defiers" (defined as individuals who would have received the "treatment" if they had been assigned to the control group but do not receive it when assigned to the program group). Unless this relationship varies across these two groups — in this example, the effect of training (X_i) on earnings (Y_i) is different for compliers than for defiers — there is no bias in the estimator β_i . However, in reality it is impossible to prove that a decision to defy is exogenous to the relationship between X_i and Y_i, so any violation of the monotonicity assumption is potentially serious.

The fifth assumption is known as the exclusion restriction. It states that any effect of the program P_i on the outcome Y_i must be mediated by the policy variable X_i for P_i to be a valid instrument for X_i. When there is an effect of P_i on Y_i that is not mediated by X_i, the instrumental variable estimator may misattribute this effect to X_i, causing the estimate of β_i to be biased. The closer program P_i and policy variable X_i are conceptually, the less likely such a violation of the exclusion restriction is likely to occur, and the less severe the bias will be when it does. (Recall that the causal inference about the policy effect β_i would be strongest if X_i itself were subject to random assignment, in which case $P_i = X_i$.)

In summary, the first three of these five assumptions are fairly easily met in most social experiments, provided that the experiment has a sufficiently strong effect on the policy variable, random assignment is carried out well, and the scale of the experiment is small enough to pre-

clude significant neighborhood or displacement effects. However, the monotonicity assumption and the exclusion criterion require considerable scrutiny if a social experiment is to be used as the basis for an instrumental variables analysis.

Key Assumptions for Identifying an IV Model with Data from a Random Assignment Design (Calculating LATE)

Instrument is Exogenous. The instrument, P_i, is randomly assigned, uncorrelated with unobserved characteristics of persons *i*, including pre-program versions of outcome Y_i.

Instrument has a meaningful effect. The instrument, P_i , must be a reliable predictor of the policy variable, X_i.

SUTVA (stable unit treatment value). The value of the policy variable, X_i, and the relationship between it and the outcome, Y_i, is not affected by variation in X and Y for others.

Monotonicity. The effect of the instrument, P_i, on the policy variable, X_i, is not less than or smaller than an effect on X_i that would otherwise occur.

Exclusion restriction. Any effect of the instrument, P_i, on the outcome, Y_i, must be mediated by the policy variable, X_i, conditional on observed characteristics.

Consequences of Defiance

To address the monotonicity assumption, it is necessary that researchers have a clear understanding of the direction of the expected program effects δ_i (of P_i on X_i) and the extent and nature of defiance among study participants in the experiment.

To illustrate the potential effect of defiance on instrumental variables estimates from a social experiment, consider the possibility of using a multigroup random assignment evaluation of welfare-to-work programs to study the effects of maternal employment on child outcomes. In the National Evaluation of Welfare-to-Work Strategies (NEWWS), funded by the U.S. Department of Health and Human Services and conducted by MDRC, sample members in three of the seven sites were randomly assigned to one of three research groups: a labor force attachment (LFA) group; a human capital development (HCD) group; or a control group. All the sample members were welfare recipients at the time of random assignment. LFA group members were offered services to accelerate their entry into employment, and HCD group members were offered services aimed at increasing their education (see Hamilton et al., 2002, for more details on

this study). Both programs produced significant positive effects on employment and earnings over time, although increases were significantly larger for the LFA program. Given these program effects, it would be tempting to use the NEWWS data to study the effects of increased employment among welfare recipients on a host of other outcomes, such as the well-being of their families or their children's education outcomes. However, using the LFA and HCD program variables as instruments for increased employment is problematic, especially for people assigned to the HCD programs, because doing so requires assuming that the expected value of program effect δ_i is positive. But many HCD program group members may have *reduced* their employment, at least initially, to participate in education and training offered through the program. In the context of the employment-based instrumental variables analysis, these participants would be considered defiers because, even though they met the program's requirements, they reduced their employment more than they would have in the control group. It is impossible to know whether employment would have similar effects on nonemployment outcomes for this group as for compliers, that is, people who increased their employment immediately rather than returning to school. Consequently, it would be preferable to limit this hypothetical instrumental variables analysis to the LFA data, where the effect δ_i is less ambiguous and defiance (in terms of the effect of the program on employment) is much less likely. Alternatively, as described later in the paper, the HCD data could be used to answer questions about the effects of maternal education on children's outcomes.

Multiple Pathways

The exclusion restriction is probably the strongest of the five assumptions and is the most difficult to verify conclusively. In one of the examples mentioned above, P_i was assignment to a training program and X_i was defined as receipt of training. In that case, P_i and X_i were closely related and most of the effect of P_i on Y_i was likely to be mediated by X_i. However, many training programs provide services other than training to their students, such as job search assistance and referrals to other education providers. Unless these services are somehow controlled for in the analysis, the benefits from these auxiliary services may be attributed to the training variable X_i and become part of the estimated effect β_i . In the case of this hypothetical training program, the consequences of this particular bias may not be severe, because it essentially redefines the treatment to include the other program components together with the training. However, when programs are more comprehensive and multifaceted, it becomes very difficult to separate the effects of one program component from the effects of another.

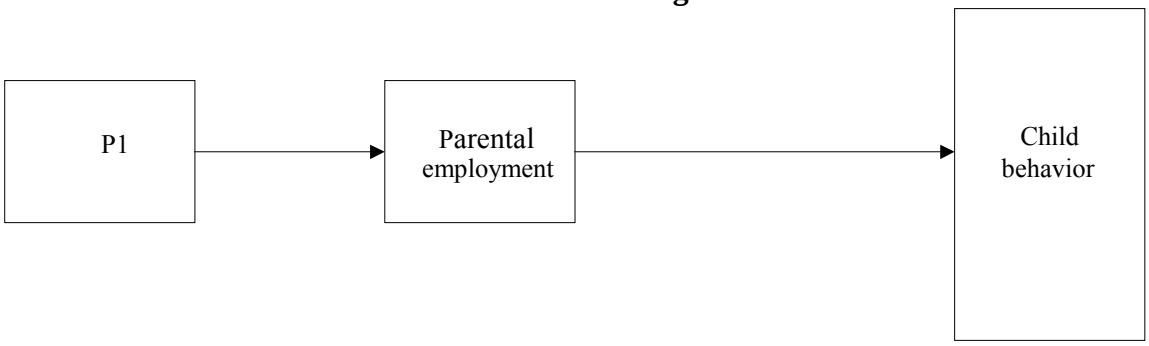
Examining More Than One Causal Path: Multiple Mediators

As just described, the strongest of the five assumptions necessary to identify an IV model is the exclusion restriction requiring that a variable P_i (e.g., assignment to the treatment group) can be used as an instrument for the effect of X_i (e.g., participation in a program) on the outcome of interest Y_i only if the relation between P_i and Y_i is fully mediated by P_i. Meeting this assumption is difficult in practice because programs often aim to affect multiple aspects of behavior and are composed of a variety of requirements, services, and incentives to achieve multiple goals (e.g., to increase employment through work requirements and reduce poverty through earnings supplements).[5] Consequently, the availability of one instrument is often insufficient to capture all of the induced behavior changes that constitute the program effect of P_i on Y_i. Under these circumstances, using a randomly assigned program variable to estimate the effect of a single X_i on Y_i may violate the exclusion restriction and will result in biased estimates of the effects of X_i on Y_i.

For example, consider a research project that uses a random assignment evaluation of a welfare-to-work program, P1, to analyze the effect of parental employment on child behavior. Is children's behavior affected by their parents' decision to seek employment? Figure 1 illustrates how such an analysis could be structured.

Figure 1

Using Program P1 to Analyze the Effect of Parental Employment on Child Well-Being



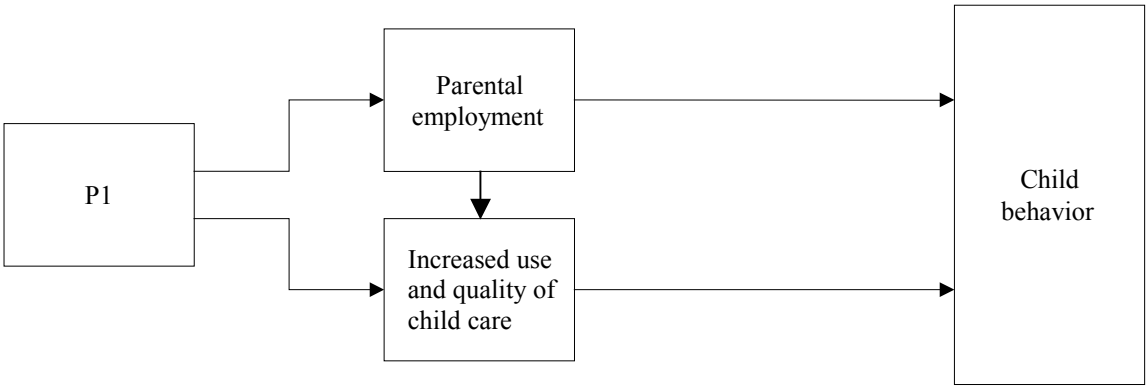
In this figure, P1 represents an employment program that offers an earnings supplement to people who work full time. It is used as an instrument for parental employment in an effort to pro-

[5]The simple ratio presented in Equation 2, $\frac{dY_i}{dX_i}$, will no longer produce a valid IV estimate.

duce an unbiased estimate of the effect of parental employment on child well-being. Through its rules and services, program P1 is expected to increase parental employment, which in turn is expected to affect child behavior. However, for this analysis to be valid, the exclusion restriction requires that parental employment be the *only* pathway through which P1 affects child behavior. And programs like P1 often provide parents with additional services, such as child care subsidies and advice on how to find good-quality child care. This aspect of program P1 could constitute a separate pathway through which P1 might affect child behavior, as illustrated in Figure 2. While some of the increased child care may be a result of the increased employment among parents (and thus captured in the total effect of employment on children), some parents may alter their use of child care even without changing their employment behavior. That is, even if the program has no effect on employment, parents in the program may use more or different types of child care than other parents. In that case, attributing all of P1's effect on child behavior to changes in parental employment is incorrect and will lead to biased estimates of the effect of parental employment on child behavior.

Figure 2

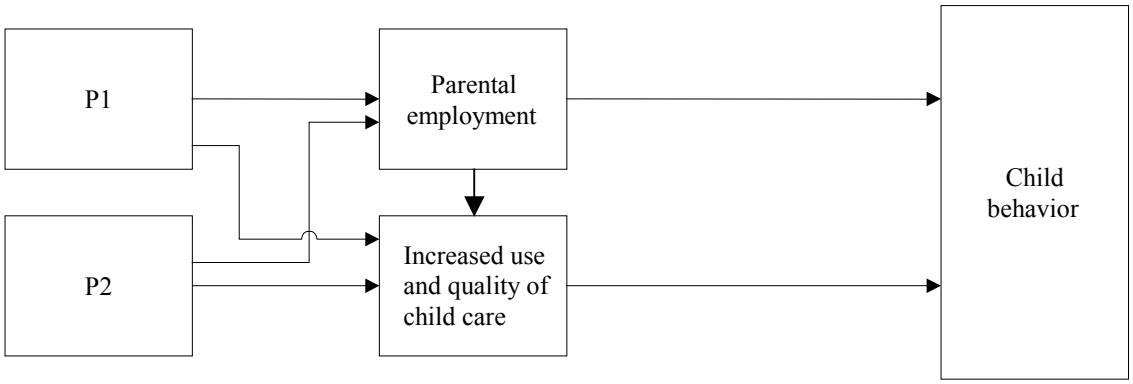
Programs Can Affect Employment and Child Care



In order to address this problem, it is necessary to introduce a second instrument, P2, as illustrated in Figure 3. As will be discussed, this could be a second randomly assigned treatment (one emphasizing employment or child care use to a different extent than P1), or it could be a different program site or welfare office.

Figure 3

Using Two Program Group Variables to Separately Identify Parental Employment and Child Care



Multiple Mediators and Identifying IV Estimates

Identifying the empirical IV model depends on satisfaction of the exclusion restriction. A model with more than one endogenous variable is "unidentified" by a single instrument and requires multiple instruments to achieve identification. More specifically, a model is unidentified if there is not enough information to estimate all of its parameters. In effect, there are an infinite number of parameters that can satisfy the conditions specified by the model. For the model to be identified, there must be at least one instrument for each endogenous variable that is being used as a predictor in the second-stage equation. If the number of exogenous instruments in the model is equal to the number of endogenous variables that need to be instrumented, the equation is "just-identified" or "exactly identified." In this case, there is just enough information to estimate the parameters needed.[6] If the number of instruments in the model exceeds the number of endogenous variables, the equation is "overidentified."[7] In the next section, we describe a number of possible approaches to creating multiple instruments using random assignment experiments to estimate the impacts of more than one policy variable (or more than one mediator).

[6]If one instrument is simply a linear combination of another, the equation would still be unidentified, even with the same number of instruments as endogenous variables. As in the case in which too few instruments are included, there are an infinite number of parameters that can satisfy the equations.

[7]One advantage to overidentified models is that they allow a test of the validity of the instruments. More specifically, a test of the overidentifying restrictions can be conducted to test whether there is any association between the instruments and the error term in the second-stage equation. Because the instruments are supposed to reflect random variation in X_i, they should not have any such association. A lack of association between the instruments and the error term provides some indication that the instruments are valid. There are several possible explanations for an association between the instruments and the error term. One is that the equation is misspecified.

Creating Multiple Instruments: Multigroup Randomized Experiments

One approach to constructing multiple instruments is to exploit a multiple group research design within a study. In a multiple group research design, subjects are randomly assigned to one of several program groups or to a control group. An example of using IV estimation with data from a study that employed a three-group research design to identify the effects of income on children's well-being is described in Box 1. The multigroup research design provides access to several valid instruments, with separate program dummies representing assignment to the first program group, the second program group, and so forth. The original one-policy variable model is thus expanded to:

$$Y_i = \alpha + \sum_{1}^{m} \beta_{im} X_{im} + \sum_{1}^{k} \beta_{ik} Z_{ik} + \varepsilon_i \,, \tag{8}$$

where X_{im} are m potentially endogenous policy variables (child care use and employment in the example above) and β_{im} are the estimated effects associated with those variables. Expressed as a set of IV models, this equation can be written as follows:

$$Y_i = \alpha + \sum_{1}^{m} \beta_{im} X_{im} + \sum_{1}^{k} \beta_{ik} Z_{ik} + \varepsilon_i \text{ , where for every } X_{im} \,, \tag{9}$$

$$X_{im} = \gamma + \sum_{1}^{s} \delta_{is} P_{is} + \mu_i \text{ , in which} \tag{10}$$

μ_i is the first-stage error term and other covariates Z_{ik} are omitted for simplicity. A necessary condition for this system of equations to be identified is that *s,* the number of independent instruments, be equal to or greater than *m,* the number of mediators X_i through which P_i affects Y_i.

Of course, even after multiple instruments are added, the exclusion restriction still applies: The only way that the instruments are assumed to affect the outcome is through the pathways that were explicitly included in the model. This assumes that any other important pathways are not omitted, either because they were not measured or because there were not enough instruments to estimate their effects. Thus, implicitly, the contribution of any of these other pathways to changes in outcome Y_i (e.g., in the example above, any family-level income effects associated with the increase in parental employment) is absorbed by the effects associated with the included pathways.

Arguably, multigroup random assignment designs are the cleanest way to produce multiple instruments with which to estimate complicated multimediator models like those introduced above, because they provide more than one truly exogenous instrument. However, it is

difficult and expensive to carry out such designs in a real-world program environment. Also, for such designs to work in the context of an instrumental variables analysis, there must be sufficient variation in the effects of the individual program variables on the mediating policy variables (i.e., in Figure 3, the effects of P1 and P2 on the two mediators must differ from one another). For example, Box 1 shows that the Minnesota Family Investment Program (MFIP) had a larger effect on employment than did an inventives-only variant of the program. However, the effects of MFIP and MFIP Incentives Only on income were relatively similar, which may be one reason why the IV estimates are not larger. Otherwise, we have, in effect, one rather than two independent instruments. All of this means that there are relatively few multigroup social policy experiments that have sufficient truly random variation in their treatments to carry out complex IV estimation procedures like those outlined above. Other approaches may be necessary — and are discussed below.

Box 1

An Empirical Example of IV Analysis with Multiple Mediators Using Data from a Multigroup Research Design

In this study, data from the Minnesota Family Investment Program (MFIP) were used to estimate the effects of income on child well-being. In MFIP, single-parent families receiving welfare were randomly assigned to one of three research groups: (1) Full MFIP, (2) MFIP Incentives Only, or (3) AFDC (the control group). Whereas under AFDC earnings reduced welfare payments dollar for dollar, families in both Full MFIP and MFIP Incentives Only were able to keep more of their welfare income as their earnings increased. In addition, families in the MFIP group were required to participate in employment and training services if they had been on welfare for 24 of the prior 36 months (or else face sanctions), while those in the Incentives Only group did not face any of these employment and training mandates. Families assigned to the AFDC group received the benefits of Minnesota's AFDC program.

(continued)

Box 1 (continued)

The first-stage equation predicting income included the two instruments (P_1 and P_2), representing assignment to each of two research groups, and a set of baseline characteristics hypothesized to affect income and employment. A similar first-stage equation predicting employment was also estimated. Because MFIP's effects on employment and income were strongest in the first year than later in the follow-up period, income and employment in the first year of the study were used as the dependent variables in the first-stage equation (Miller et al., 2000).

The results of the first-stage equations are presented in Table 1.1 below. As is evident from the table, the dummy variables representing Full MFIP and MFIP Incentives Only were associated with employment and income, a necessary condition for the IV strategy. A test of the effects of the instruments suggests that these variables are strong predictors of both employment ($F = 13.38$, $p < .001$) and income in the first year ($F = 14.65$, $p < .001$).

Table 1.1

The Effects of MFIP on Employment and Income: First-Stage Regression Results for IV Model

	Year One Employment	Year One Income	Three Year Employment	Three Year Income
Full MFIP	0.20***	1.40***	0.16***	1.23***
	(0.04)	(0.29)	(0.03)	(0.36)
MFIP Incentives Only	0.09**	1.32***	0.08***	1.10***
	(0.04)	(0.29)	(0.03)	(0.37)
F value	13.38***	14.65***	15.13***	6.89***
Sample Size= 879				

SOURCE: MDRC calculations using MFIP administrative and baseline survey data.

NOTES: Standard errors in parentheses.

The sample includes long-term welfare recipients randomly assigned from April 1, 1994, to October 31, 1994, excluding the small percentage who were receiving only Food Stamps at random assignment.

Income is measured in thousands of dollars, in the first year after random assignment and on average over the three-year follow-up. Employment is measured as ever employed, in the first year after random assignment and on average over the three-year follow-up.

The regressions also include the following covariates measured at baseline: black, other racial/ethnic minority, mother was a teen at child's birth, number of children in the family, presence of a child age 6 or less, mother had no high school degree or equivalent, mother never married, mother on welfare 5 or more years, earnings in the prior year, and indicators for the quarter of random assignment.

Two-tailed significance levels are indicated as: * = 10 percent; ** = 5 percent; *** = 1 percent.

(continued)

Box 1 (continued)

The second-stage equation used predicted income and employment (e.g., $\hat{Z}_{1i}$ and $\hat{Z}_{2i}$) along with the same set of baseline characteristics (excluding the instruments) to predict the child outcomes.

The results of this second-stage equation are presented in Table 1.2. For comparison, the results of analogous OLS estimation methods are also provided. (In the OLS models, the same covariates are used as in the second-stage IV estimate, and income and employment are both included in the equation, rather than the predicted value of income and employment as in the IV models.) In the OLS models, small, insignificant effects of income are found. However, in the IV models, significant positive effects of income are found, predicting engagement in school and positive social behavior. The effects on the other two variables are in the expected direction (favorable) but are not statistically significant. In none of the models did employment have a significant effect (although, interestingly, the coefficients on the employment measures are always opposite to those of the income measures).

(continued)

Box 1 (continued)

Table 1.2

OLS and IV Estimates of the Effects of Employment and Income on Children's School and Behavioral Outcomes

	OLS	IV Model 1	IV Model 2
	Effects of Income		
School Achievement	-0.02	0.16	0.20
(mean = 4.06, sd = 1.10)	(0.01)	(0.14)	(0.20)
		H: p = .02	*H: p = .02*
School Engagement	-0.01	0.47*	0.59
(mean = 10.10, sd = 1.82)	(0.02)	(0.27)	(0.42)
		H: p = .01	*H: p = .01*
Behavior Problems	-0.16	-1.57	-1.94
(mean = 11.69, sd = 9.20)	(0.12)	(1.23)	(1.72)
		H: p = .11	*H: p = .10*
Positive Behavior	0.17	11.04*	14.01
(mean = 196.16, sd = 37.5)	(0.49)	(6.18)	(9.32)
		H: p = .06	*H: p = .06*
	Effects of Employment		
School Achievement	-0.02	-0.17	-0.36
(mean = 4.06, sd = 1.10)	(0.09)	(1.08)	(1.66)
		H: p = .02	*H: p = .02*
School Engagement	0.13	-1.05	-1.87
(mean = 10.10, sd = 1.82)	(0.16)	(1.86)	(3.17)
		H: p = .01	*H: p = .01*
Behavior Problems	0.45	1.99	4.32
(mean = 11.69, sd = 9.20)	(0.80)	(8.30)	(12.87)
		H: p = .11	*H: p = .10*
Positive Behavior	-3.19	-65.05	-94.62
(mean = 196.16, sd = 37.5)	(3.24)	(41.63)	(69.74)
		H: p = .06	*H: p = .06*
Sample Size= 879			

(continued)

Box 1 (continued)

SOURCE: MDRC calculations using MFIP administrative, child survey and baseline survey data.
NOTES: Standard errors in parentheses. "H" indicates Hausman test (p values of F test are indicated).

The sample includes long-term welfare recipients randomly assigned from April 1, 1994, to October 31, 1994, excluding the small percentage who were receiving only Food Stamps at random assignment.

Income is measured in thousands of dollars, in the first year after random assignment (model 1) and on average over the three-year follow-up (model 2). Employment is measured as ever employed, in the first year after random assignment (model 1) and on average over the three-year follow-up (model 2).

The regressions also include the following covariates measured at baseline: black, other racial/ethnic minority, mother was a teen at child's birth, number of children in the family, presence of a child age 6 or less, mother had no high school degree or equivalent, mother never married, mother on welfare 5 or more years, earnings in the prior year, and indicators for the quarter of random assignment.

Two-tailed significance levels are indicated as: * = 10 percent; ** = 5 percent; *** = 1 percent.

Creating Multiple Instruments: Multisite Randomized Experiments

An alternative approach that can be implemented post hoc is to exploit the variability that occurs due to the implementation of comparable experiments across multiple sites or offices. It is possible to create more than one instrument by interacting the random assignment treatment variable with a variable representing each of the sites. The result is an analysis in which variation in the implementation of the program is used to identify the various pathways through which the program affected the outcome. This approach works best when there is a fairly large number of sites or offices and program implementation was varied either deliberately to produce variation in program variables P_i or naturally for reasons exogenous to the policy variables X_i and the outcome variables Y_i. Such exogeneity of the variation in P_i across sites is essential to safeguard the validity of the instrumental variables analysis.

Bos and Granger (2000) provide an example using a multisite approach to estimate the effect of early day care use on the school readiness of children born to teen mothers. Using data from the 16-site New Chance Demonstration (Quint et al., 1997), this paper exploits variation across the sites in program effects on day care use and other possible mediators of program effects on child outcomes to disentangle effects of different aspects of children's day care experiences on the child outcomes. Another example using data from a multisite and multigroup research design to estimate the effects of maternal education on children's cognitive outcomes is described in detail in Box 2. Note that the data from this study were earlier presented as violating the monotonicity assumption when used to estimate the effects of maternal employment on children's outcomes because welfare recipients randomly assigned to the HCD programs may have

Box 2

An Empirical Example of IV Analyses with Multiple Mediators Using Data from a Multisite and Multigroup Research Design

This study described here examined the effects of parents' participation in schooling using data from the National Evaluation of Welfare-to-Work Strategies (NEWWS; Magnusson and McGroder, 2002). The evaluation of NEWWS included six programs evaluated across three sites (Atlanta, Georgia; Grand Rapids, Michigan; and Riverside, California) that operated in the early to mid 1990s under the federal Job Opportunities and Basic Skills Training (JOBS) Program, which preceded the current welfare system, Temporary Assistance for Needy Families (TANF). The primary objective of these programs, like TANF, was to reduce single parents' welfare use and increase their employment. In one condition, single-parent welfare recipients were assigned to a program that required most participants to look for work immediately, usually by attending a "job club" that lasted one to three weeks (this condition was termed labor force attachment, or LFA). In the other condition, participants were placed in education and training programs (usually adult basic education or vocational training) to increase their knowledge and skills before they attempted to move into employment (this condition was termed human capital development, or HCD).

Each of the three sites operated *both* an LFA program stressing job search as a first activity and an HCD program stressing basic education as a first activity, single-parent welfare recipients were randomly assigned to one of these program groups or to a control group. The program groups were required to participate in basic education or employment-related activities (depending on the group) as a condition of receiving welfare. Families who failed to meet the participation requirements could receive sanctions, that is, have their welfare grants reduced.

IV models were estimated to assess what effect parents' participation in schooling had on children's cognitive test scores. Interactions between program and site were used as instruments to estimate the effects of two endogenous variables, participation in employment and participation in educational activities, on test scores assessing children's school readiness two years after parents' random assignment to the programs.

(continued)

Box 2 (continued)

The first-stage IV results are presented in Table 2.1 below. They indicate that the HCD programs in the three sites all significantly increased parents' participation in educational activities. The LFA programs in the three sites all significantly increased parents' participation in employment. And three of the programs — Atlanta LFA, Grand Rapids LFA, and Riverside HCD — significantly increased *both* education and employment. Note that the effects across the six programs are different with respect to their effects on education and employment. This variation is critical to identifying the second-stage effects in the IV model using predicted values of education and employment as independent variables in models predicting child test scores.

Table 2.1

First-Stage IV Coefficients, F-statistics, and R-squares (Standard Errors in Parentheses)

Instruments	Months of Education		Quarters of Employment	
Atlanta HCD	2.36	***	.25	
	(.34)		(.17)	
Atlanta LFA	.60	*	.43	***
	(.34)		(.17)	
Grand Rapids HCD	.96	*	.00	
	(.50)		(.25)	
Grand Rapids LFA	-.98	*	.96	***
	(.50)		(.25)	
Riverside HCD	2.94	***	.68	***
	(.43)		(.21)	
Riverside LFA	-.36		1.22	***
	(.44)		(.22)	
F-statistic for instruments	20.90	***	9.63	***
Full model R-square	.17	***	.21	***
Increase in R-square associated with instruments	.040	***	.015	***

NOTES: Two-tailed significance levels are indicated as: * = 10 percent; ** = 5 percent; *** = 1 percent.

Covariates were included for: educational attainment and participation at baseline, prior earnings, prior welfare receipt, numeracy, literacy, depressive symptoms, mother's and focal child's age, number of baseline risk factors, family barriers to employment, race, marital status, number of children, an index of one's sense of control over one's life, sources of social support, and child gender.

Box 2 (continued)

The results of the second-stage model using predicted employment and education from the first-stage equation as a predictor of children's test scores are presented in Table 2.2. Two models were estimated in the second-stage equation. In the first model, only the predicted value of education was used as an independent variable. In the second model, both the predicted value of education and the predicted value of employment were used as independent variables. In both models, participation in education had a positive, significant effect on children's test scores. The effects of employment were significant in comparative OLS models, but not in the IV model.

Table 2.2

OLS and IV Estimates of Months in Educational Activities on Children's Raw Bracken School Readiness Composite Scores (Standard Errors in Parentheses)

	Model 1: Bracken		Model 2: Bracken	
Independent variables	OLS	IV	OLS	IV
Months in education	.089 ***	.305 *	.098 ***	.311 *
	(.035)	(.168)	(.035)	(.169)
Quarters of employment			.134 *	.671
			(.070)	(.493)

NOTES: Two-tailed significance levels are indicated as: * = 10 percent; ** = 5 percent; *** = 1 percent.

Covariates were included for: educational attainment and participation at baseline, prior earnings, prior welfare receipt, numeracy, literacy, depressive symptoms, mother's and focal child's age, number of baseline risk factors, family barriers to employment, race, marital status, number of children, locus of control, sources of social support, and child gender.

initially reduced their employment in order to pursue more education. The monotonocity assumption is not violated, however, when these data are used to estimate the effects of maternal education on children's outcomes. One way to assess bias due to the violation of monotonicity in this case is to compare the IV estimates of the effects of maternal education using data from the HCD programs with the IV estimates on the effects of maternal education from the LFA programs.

Creating Multiple Instruments: Subgroups from Randomized Experiments

A similar approach can be implemented in which variation in the responses of particular subpopulations to a program P_i is used to create multiple instrumental variables. In this case, the random assignment treatment variable is interacted with one or more exogenous baseline characteristics, such as age or gender. Thus, the baseline characteristic serves as a covariate in both equations, and the interaction of the program variable and the baseline characteristic serves as one of the instruments. In equation form, this can be written as:

$$Y_i = \alpha + \sum_{1}^{m} \beta_{im} X_{im} + \sum_{1}^{k} \beta_{ik} Z_{ik} + \varepsilon_i \text{ , where for every } X_{im} \tag{11}$$

$$X_{im} = \gamma + \sum_{1}^{s} \delta_{is} P_i Z_{is} + \sum_{1}^{s} \nu_{is} Z_{is} + \mu_i \text{ ,} \tag{12}$$

in which Z_{is} is a series of exogenous baseline variables, $s < k$, and $s \geq m$.

Conceptually and technically, this approach is identical to the use of different sites as variables Z_{is}. In practice, interacting the program variable with demographic characteristics or other baseline variables as well as by site may be problematic. Variation in program effects across different subgroups in the same location or across locations may not be truly exogenous to a measure of child well-being. Selection — that is, something unique about the subgroups or sites that drives program effects — could account for part of this variation, which would undermine the IV estimates' validity. It is necessary to ensure that the relationships between the endogenous variables X_i and the outcome Y_i are also not significantly moderated by the exogenous baseline characteristics Z_{is} or by site. In other words, the effect of X_i on Y_i must be the same across different levels of Z_{is} (for example, if instruments were created by interacting the program group variable with child age or child gender, the relationship between child care use and children's school readiness must be the same across child age or child gender).

Creating Multiple Instruments: Pooling Data from Multiple Experiments

A final alternative for constructing multiple instruments is to construct a pooled data set or a data set that combines information from multiple random assignment experiments. Pooling is only possible if the type and quality of the data as well as the outcomes of interest are comparable across the pooled studies. The pooled data, as described in Box 3, offer the benefits of having an instrument for each respective random assignment study. These multiple instruments can be used for either of two purposes: (1) to estimate multiple mediators or (2) to increase the precision of

estimates for a single mediator, a topic that we turn to in the next section. Getting comparable information from enough random assignment experiments is no small feat, and it can often take years before enough studies have been conducted for such an effort to be productive.

Box 3

An Empirical Example Using Pooled Data from Multiple Random Assignment Experiments to Estimate the Effects of Income, Employment, and Child Care on Children's Well-Being

In this example, we attempt to answer questions about the effects of income, employment, and child care on children's well-being using a pooled data set of experimental studies of welfare and work programs. Each study evaluated a program using a random assignment design, and comparable data on families and children were collected across studies.

The primary equation of interest is:

$$Y_i = \alpha + Z_i\beta + E_i\beta_E + I_i\beta I + F_i\beta_F + \varepsilon_i ,$$

where *i* represents each child, F is a measure of participation in formal child care, E is a measure of parents' employment, and I is a measure of family income. The outcome variable is a measure of children's cognitive functioning. The Z's are a variety of controls or covariates predicted to affect children's cognitive functioning, such as age and education of the mother and number of siblings, and the error term is represented by ε. Using OLS techniques, the effects of F, I, or E on Y could be biased if they are correlated with the error term. Instrumental variables models can control for such biases.

In this case, the first stage in estimating such a model would require estimating three models that look something like:

$$E_i = \alpha_o + Z_i\beta + P_i\gamma + \mu_i$$

$$I_i = \alpha_o + Z_i\beta + PP_i\gamma + \eta_i$$

$$F_i = \alpha_o + Z_i\beta + PPP_i\gamma + \delta_i$$

These first three equations derive $\hat{E}, \hat{I}$, and $\hat{F}$. These predicted measures of employment, income, and formal child care replace the actual measures in the first equation displayed in this box. The resulting estimates of the effects of formal child care, income, and employment on children's cognitive functioning will be free of bias if the IV assumptions noted above hold.

(continued)

Box 3 (continued)

The Ps in these equations represent the instruments used to identify the first-stage equations. A minimum of three instruments, one for each model, are needed to estimate this model. The pooled data from multiple welfare and work programs offer a number of possibilities. The challenge is to construct a set or sets of instruments that will (1) reliably predict the outcome of interest (employment, income, or formal care) and (2) reliably distinguish the prediction of one outcome from the prediction of another outcome (e.g., do a better job at predicting employment than predicting formal care).

Three instruments that represent differing policy approaches can be used to identify the above equations. Specifically, some of the studies evaluated programs with expanded child care resources that affected parents' use of formal as opposed to informal care. Some of the studies increased employment through mandatory employment services but did not increase their income (welfare recipients in these programs traded their welfare checks for earnings). Finally, some of the studies evaluated programs that provided financial incentives to work: These increased both employment and income. The logic here is that particular policy approaches in these welfare and work programs had unique influences on certain outcomes. In other words, the programs did not all affect formal care, employment, and income in the same way.

All of these approaches (multiple treatment groups within a study, interactions with site, interactions with subgroup variables, and pooling data across studies) can be combined to create a sufficiently large number of instruments to conduct an instrumental variables analysis with multiple mediators. This can be useful in the case of additional mediating variables, as well as for the purposes of verifying the validity of the instruments (by creating overidentified models; see footnote 8).

Estimation Issues: The Problem of "Weak" Instruments

The prior sections provided a general framework for understanding the policy questions that could be answered by instrumental variables analysis and the necessary assumptions for identifying causal effects to answer such questions, including the use of multiple instruments to identify multiple mediators. Even when all of the assumptions described above are met and an appropriate IV estimator is identified, actual estimation of instrumental variables poses a set of new issues concerning both the validity and reliability of the estimates. As discussed earlier, to obtain consistent and reliable instrumental variables estimates, a good instrument P_i must be highly correlated with the policy variable X_i. In the following section, we will discuss some of

the drawbacks of using "weak" instruments (i.e., instruments that do not have strong correlations with the policy variable).

There are several risks in interpreting IV estimates obtained using a weak instrument or a set of weak instruments. First, there is the risk of having large errors on the IV estimates in the second stage of the procedure, which would make the estimates unreliable. Second, weak instruments can produce IV estimates that are vulnerable to bias due to chance correlations between the error terms in the different stages of the IV procedure. We discuss both of these risks in more detail below.

The Cause of Weak Instruments: Weak Program Effects

While the use of a randomly assigned program variable as an instrument avoids the problem of correlation with omitted variable(s), the possibility that a randomly assigned program variable is a weak instrument is real. Even if a program has its intended effect on a policy variable (e.g., increases vocational training or employment), assignment to the program group may not be the most important predictor of the policy variable relative to other potential predictors. For example, the effect of a program that seeks to increase employment may be small relative to other predictors of current employment, such as prior employment experience, educational background, or current family circumstances. This is particularly true when random assignment to treatment is used to predict variation in policy variables X_i that are not primary targets of the experiment. For example, when training or employment programs P_i increase the use of child care for parents with young children, random assignment to such programs can be used as an instrument to predict child care use (as long as a separate instrument is used to predict employment, as described in the section on multiple mediators), but the strength of the relationship and the relevance of the instrument may be limited because directly encouraging child care use was not the original intent of the program. Consequently, the program variable P_i is likely to be a better predictor of variation in employment than of variation in child care use.

This problem is compounded when predicting multiple mediators. In this case, obtaining reliable estimates will depend on strong program effects on each of the different outcomes *as well as* variation in program effects on different outcomes across the instruments. For example, if a pooled data set is composed of data from a set of random assignment studies of employment programs and all of the studies increased employment and did not have any effects on income or child care, then the program variables (even though there is more than one) cannot be used as multiple instruments to predict income or child care.

One Consequence of Weak Instruments: Unreliable IV Estimates

To better understand the implications of having an irrelevant or weak instrument, recall Equation 2. If P_i is not relevant to X_i or if P_i is a very weak predictor of X_i, then $\frac{dX_i}{dP_i}$ — the independent share of the variation in X_i that is due to P_i — will be negligible relative to the total variation in X_i and not sufficient to purge the relationship between policy variable X_i and outcome Y_i of spurious covariation. This leads to invalid, inconsistent, or unreliable IV estimates, as evidenced by large standard errors on the IV estimates in the second stage of the procedure.

An example of the implications of weak program effects is described in Box 2. In this case, even though the program was designed to increase employment and income (the two endogenous variables that are being estimated in a first stage), program effects on employment and income were stronger during the first year relative to the average effects during the three-year follow-up period. As a result, the IV estimates of the effects of income on child well-being are more precisely estimated using the measure of income during the first year. The IV estimates of income averaged over the three years were similar in magnitude but were less precise (i.e., the estimate on income had a much larger standard error). This can be seen by comparing the results of model 1 with those of model 2 in Table 2.2.

A Second Consequence of Weak Instruments: Biased IV Estimates

In finite samples, even good instruments cannot ensure that estimates are unbiased.[8] Consider a finite sample with an instrument P_i, an outcome variable Y_i, and a mediator X_i. There is no true effect of X_i on Y_i, but both X_i and Y_i are correlated with an unmeasured variable Z_i. The instrument P_i is a randomly created program variable and is uncorrelated with Z_i by construction. However, in a finite sample, P_i may be correlated with Z_i by chance. Through this correlation, and in this finite sample, the instrument P_i will reintroduce a spurious effect of X_i on Y_i (referred to as "finite sample bias"). Intuitively, finite sample bias arises because IV estimates rely on the preciseness of the estimates of the first-stage coefficient rather than the coefficient's actual value. Even if there were no relationship between P_i and X_i, the estimates of the relationship

[8] An unbiased estimate means that the estimate has a sampling distribution centered on the parameter of interest in a sample of *any* size. Because IV estimates are based on a ratio of random quantities, the expectation of such a ratio does not necessarily have a simple form. A consistent estimate means that the parameter converges to the population parameter as the sample size grows.

between P_i and X_i would not be zero in any finite sample. A number of researchers have examined the finite sample properties of IV (e.g., Sawa, 1996; Staiger and Stock, 1994).[9]

Using large samples is one way to minimize this bias as well as to have more precise estimates (i.e., estimates with smaller standard errors). However, when the relationship between instrument P_i and policy variable X_i is weak, such finite sample bias can be significant even when samples are very large. Bound et al. (1995) demonstrate that the typical method for minimizing finite sample bias, increasing the sample size, does not solve this problem, especially for estimates obtained from large cross-sectional samples. In other words, large data sets do not necessarily insulate IV estimates from finite sample bias caused by weak instruments.

Addressing the Problem of Weak Instruments

Given the risks of using a weak instrument, one useful guideline in pursuing IV estimation is to perform a close examination of the characteristics of the first-stage equation or equations. The stronger the relationship between P_i and X_i (i.e., the greater the partial R^2 on the excluded instruments), the lower the likelihood that weak instruments will bias the IV estimates. Testing estimates with alternative instruments is one potential approach to assessing the robustness of IV estimates or creating bounds for these estimates similar to confidence intervals on traditional OLS estimates. However, when randomly assigned program variables are used as instruments, there usually is no ready supply of such alternative instruments.

There are other techniques to address the problem of finite sample bias in cases where the available instrument or instruments are fairly weak and no good alternatives are available. The key to making such techniques work is to break the link between spurious correlations between P_i and X_i and similar correlations between P_i and Y_i. One such technique relies on split sample estimates (see, e.g., Angrist and Krueger, 1995). In this approach, two samples are drawn from a single population, ideally at random, and the first and second stages of the instrumental variables analysis are carried out separately on those two independent samples. That is, the first sample is used to estimate the relationship between P_i and X_i. Using the regression coefficients from this analysis, $\hat{X}_j$, a predicted value of X_i is estimated in the second sample, and the outcome Y_j is regressed on $\hat{X}_j$ in the second stage of the analysis, resulting in β_j, an IV estimate of the effect of X on Y. Angrist and Krueger (1994) show that sampling variation tends to bias β_j toward zero, as P_i is never going to predict X_j as well as X_i and will therefore introduce additional random error into the estimation of the relationship between X_j and Y_j. However, this is usually preferable to a situation in which β_j is biased towards the original biased OLS esti-

[9]Examining the partial R^2 and the F-statistic on the instruments in the first-stage regression helps gauge the potential finite sample bias of IV relative to OLS (Bound et al., 1995). In fact, there are useful guidelines for assessing whether or not the explanatory power of the instruments in the first stage is adequate.

mator, as is the case when finite sample bias is present. Currie and Yelowitz (1997) apply this technique using data in a paper that explores the relationship between living in public housing and child outcomes.

Ultimately, given any finite sample, choosing instruments requires striking a delicate balance between efficiency and bias. The best instruments are those that obtain IV estimates that are asymptotically efficient *and* have small finite sample bias. Though randomized experiments can provide a valid instrument that may yield consistent IV estimates, a different choice of instrument will yield different estimates in any finite sample. Thus, there are risks to changing (or increasing the number of) instruments in finite samples without adhering to some empirical (or quantifiable) standard about the quality of the instrument. It is straightforward to determine the quality of an instrument empirically in the case when only one endogenous variable is being considered, but the risks associated with using weak instruments escalates when more than one instrument is needed to identify an IV model.

Discussion and Conclusions

For years, policymakers and researchers have grappled with developing empirical techniques to better understand important relationships between economic behavior and self-sufficiency and family or child well-being. Experiments offer the kind of exogenous variation that can help to empirically identify these relationships. There are experiments that occur naturally and those that we can create, and both have their virtues as well as problems. We argue that applying instrumental variables techniques to data from random assignment designs can be a powerful method for answering important policy questions. The goal in this chapter was to make the understanding and application of instrumental variables techniques accessible to a wide range of policymakers and researchers.

The availability of data from numerous recent random assignment studies (e.g., of welfare and employment programs) provides a unique opportunity for researchers to dig into the black box and tackle difficult questions about *how* programs affect outcomes. IV estimates do not on their own answer all policy-relevant questions but can provide policy-relevant estimates (i.e., local average treatment effects).

The application of instrumental variables, however, should not be foolhardy. One of the benefits of using a randomly assigned treatment as an instrument is that many of the key assumptions of instrumental variables techniques are identified. Nonetheless, two of these assumptions — monotonicity and the exclusion restriction — must be carefully checked and addressed. As the empirical examples in this paper show, it is frequently the case that more than one instrument is needed because programs have multiple goals and are likely to directly affect multiple outcomes, and it is almost never the case that even multigroup research design pro-

grams produce substantially different effects on related outcomes (e.g., it is unusual to have an experiment with a three-group research design in which one of the programs produces substantially different effects on a particular outcome relative to the other program). In addition, as shown in Box 2, potential violation of the monotonicity assumption implies that researchers need to consider carefully whether the random assignment design and resulting program effects are appropriate for answering the policy or research question of interest.

While substantial progress has been made in understanding the assumptions underlying instrumental variables estimation, these assumptions are best understood under specific conditions, that is, when there is one policy variable of interest.[10] Estimating multiple mediating pathways remains a key methodological challenge in instrumental variables analyses involving the currently available data from randomized experiments. Very few social policy experiments to date have included multiple randomly created "treatments," and many that do were not designed to produce significant variation in a range of important policy variables. This leaves researchers with little choice but to create multiple instruments based on variation of program effects across sites and subgroups. This approach is promising in some cases but can reintroduce bias that the instrumental variables procedure was designed to remove. It also relies on variation in program implementation across sites or subgroups, variation that often is not substantial enough to produce reliable estimates free of finite sample bias.

In addition to expanding opportunities to learn from current data, instrumental variables estimation highlights important future research opportunities. Randomized experiments could be designed specifically to produce significant variation in key policy variables in such a way that unbiased estimates of the effects of those variables on key outcomes could be obtained. The focus in such studies would not be on the program effects per se but on the secondary effects of the mediators on the outcome. Combining the tools of instrumental variables and random assignment in such designs from the outset could dramatically improve the quality of instrumental variables analyses based on random assignment. One hypothetical example is described in Box 4. Policy researchers and program operators should work together to identify opportunities for such studies, which would be more useful than traditional random assignment experiments and more valid than traditional nonexperimental approaches to policy research.

[10] Another condition in which IV assumptions are well understood (which we do not discuss here) is when the underlying outcome of interest is linear.

Box 4

Example of a Possible Future Experiment that, with IV, Can Measure Causal Relationships

A key welfare policy question concerns the extent to which transitional Medicaid benefits affect the well-being of welfare leavers and their ability to remain off welfare. These benefits are expensive to administer, and take-up is generally low. Policymakers are concerned that families who do not use transitional Medicaid are less likely to use preventive medical care and more likely to end up in emergency rooms. It is difficult to assess the effects of transitional Medicaid on families, because take-up of these benefits is selective. More advantaged individuals are more likely to know about the program, whereas less healthy individuals are more likely to find out about it because of an illness or hospital visit.

A random assignment study with an IV component could be designed to assess the effectiveness of these transitional benefits. Such a study would use a so-called "encouragement design" in which we would not change or extend Medicaid benefits (which would be very expensive) but rather seek to increase awareness of existing program services among those already eligible, keeping track of those who were randomly targeted for additional information and help in accessing benefits. Using instrumental variables, it would be straightforward to estimate the effect of the transitional Medicaid services from the experimental effect of the encouragement of its use (provided that the effect is sufficiently large).

The actual treatment in a study like this could have a tiered structure. (Using multiple tiers would help in the encouragement design by providing multiple potential instruments.) For example, Level I could be to simply send a random subset of eligible people in a county a letter as soon as they become eligible, providing a phone number and explaining in multiple languages what the program is like and how it could help. Level II would be to call or visit eligible families in order to take a more active role in making sure they use the service. Level III would be to add an ombudsman type of person, who would assist with eligibility determination, advocate with doctors and dentists over acceptance of benefits, and help resolve other administrative problems that come up in use of benefits. Obviously, the cost of administering these treatments would increase with their extensiveness, but the cost would never include the prohibitive expense of actually providing benefits.

References

Angrist, Joshua. Forthcoming, 2002. "How Do Sex Ratios Affect Marriage and Labor Markets? Evidence from America's Second Generation." *Quarterly Journal of Economics.*

Angrist, Joshua, with Guido Imbens and Alan Krueger. 1999. "Jackknife Instrumental Variables Estimation." *Journal of Applied Econometics, January-February.*

Angrist, Joshua, Guido Imbens and Don Rubin. 1996. "Identification of Casual Effects Using Instrumental Variables." JASA Applications invited paper, with comments and authors' rejoinder. *Journal of the American Statistical Association* (91).

Angrist, Joshua, and Alan Krueger. (Forthcoming 2001). "Instrumental Variables and the Search for Identification." *Journal of Economic Perspectives.*

Angrist, Joshua, and Alan Krueger. 1995. "Split-Sample Instrumental Variables Estimates of the Return to Schooling." *Journal of Business and Economic Statistics.*

Angrist, Joshua, and Guido Imbens. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica.*

Angrist, Joshua, and Alan Krueger. 1992. "The Effects of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples." *Journal of the American Statistical Association*.

Angrist, Joshua, and Alan Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics,* November.

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

Bos, Johannes, Susan Scrivener, Jason Snipes and Gayle Hamilton. 2001. *Improving Basic Skills: The Effects of Adult Education in Welfare-to-Work Programs*. Washington, DC: U.S. Department of Education, Office of the Under Secretary and Office of Vocational and Adult Education; and U.S. Department of Health and Human Services, Administration for Children and Families and Office of the Assistant Secretary for Planning and Evaluation.

Bos, Johannes, and Robert Granger. 2000. *Estimating the Effects of Day Care Use on Children's School Readiness: Evidence from the New Chance Demonstration.* Manuscript. New York, NY: Manpower Demonstration Research Corporation.

Bound, John, David Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variables Is Weak." *Journal of the American Statistical Association,* 90(430): 443-450.

Buse, Alan. 1992. "The Bias of Instrumental Variables Estimators." *Econometrica,* 60: 173-80.

Card, David, 1999. "The Casual Effect of Education on Earnings." In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics* Volume 3. Amsterdam: Elsevier.

Card, David, and Alan Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania." *American Economic Review,* 84.

Cameron, S., and James J. Heckman. 1998. "Life Cycle Schooling and Educational Selectivity: Models and Choice." *Journal of Political Economy*, 106 (2).

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Currie, Janet, and Aaron Yelowitz. 1997. "Are Public Housing Projects Good for Kids?" National Bureau of Economic Research Working Paper No. W6305. Cambridge, MA.

Hamilton, Gayle, Stephen Freedman, Lisa A. Gennetian, Charles Michalopoulos, Johanna Walter, Diana Adams-Ciardullo, and Anna Gassman-Pines. 2002. *How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs.* New York: Manpower Demonstration Research Corporation.

Heckman, James J., with Edward Vytlacil. 1998. "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return Is Correlated with Schooling." *Journal of Human Resources,* Fall 1998.

Heckman, James J. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions." *Journal of Human Resources.* Summer.

Heckman, James J. 1996a. "Randomization as an Instrumental Variable." *Review of Economics and Statistics*, Vol. LXXVIII, pp. 336-341.

Heckman, James J. 1996b. "On Air: Identification of Casual Effects Using Instrumental Variables." *Journal of The American Statistical Association.*

Hotz, Joseph, Susan W. McElroy, and Seth Sanders. 1999. "Teenage Childbearing and Its Lifecycle Consequences: Exploiting a Very Natural Experiment." National Bureau of Economic Research Working Paper No. W7397. Cambridge, MA.

Hoxby, Caroline. 2000. "Does Competition Among Public Schools Benefit Students and Taxpayers?" *American Economic Review* 90(5): 1209-38.

Magnusson, Katherine A., and Sharon McGroder. 2002. *The Effects of Increasing Welfare Mothers' Education on Their Young Children's Academic Problems and School Readiness.* Illinois: Northwestern University.

Miller, Cynthia, Virginia Knox, Lisa Gennetian, Martey Dodoo, Johanna Hunter, and Cindy Redcross. 2000. *Reforming Welfare and Rewarding Work: Final Report on the Minnesota Family Investment Program: Volume 1: Effects on Adults.* New York: MDRC.

Morris, Pamela, and Lisa Gennetian. 2002. "Identifying the Effects of Income on Children Using Experimental Data." Under revision *Journal of Marriage and the Family.*

Morris, Pamela, Aletha Huston, Greg Duncan, Danielle Crosby, Johannes Bos. 2001. *How Welfare and Work Policies Affect Children: A Synthesis of Research.* New York: Manpower Demonstration Research Corporation.

Orr, Larry. 1999. *Social Experiments: Evaluating Public Programs with Experimental Methods.* Thousand Oaks, California: Sage.

Quint, Janet C., Johannes M. Bos, and Denise F. Polit (1997). *New Chance: Final Report on a Comprehensive Program for Young Mothers in Poverty and Their Children.* New York: Manpower Demonstration Research Corporation.

Robins, Philip, and David Greenberg. 1986. "The Changing Role of Social Experimentation in Policy Analysis" (with P. Robins). *Journal of Policy Analysis and Management*, Winter 1986.

Robins, James M., and Sander Greenland. 1996. Comment on "Identification of Casual Effects Using Instrumental Variables" by J. Angrist, G. Imbens, and A. Krueger in a JASA Applications invited paper. *Journal of the American Statistical Association,* 91.

Sawa, T. 1969. "The Exact Sampling Distribution of Ordinary Least Squares and Two-Stage Least Squares Estimators." *Journal of the American Statistical Association*, 64: 923-937.

Staiger, Douglas, and James H. Stock, 1994. Instrumental Variables Regression with Weak Instruments. Technical Working Paper 151. Cambridge, MA: National Bureau of Economic Research.

Tyler, John, Richard Murnane, and John Willet. 2001. "Who Benefits from a GED? Evidence for Females from High School and Beyond." Working Paper No. 2001-35. Providence, RI: Brown University.

Wascher, William, and David Neumark. 1994. "Reply to Card, Katz and Krueger: Employment Effects of Minimum and Subminimum Wage." *Industrial Labor Relations Review,* 417-512.

Wascher, William, and David Neumark. 1992. "Evidence on Employment Effects of Minimum and Subminimum Wage: Panel Data on State Minimum Wage Laws." *Industrial Labor Relations Review,* 58-81.