Natural and Quasi-Experiments in Economics

Bruce D. MEYER
Department of Economics, Northwestern University, Evanston, IL   60208

Using research designs patterned after randomized experiments, many recent economic studies examine outcome measures for treatment groups and comparison groups that are not randomly assigned. By using variation in explanatory variables generated by changes in state laws, government draft mechanisms, or other means, these studies obtain variation that is readily examined and is plausibly exogenous. This article describes the advantages of these studies and suggests how they can be improved. It also provides aids in judging the validity of inferences that they draw. Design complications such as multiple treatment and comparison groups and multiple preintervention or postintervention observations are advocated.

KEY WORDS: Comparison groups; Control groups; Difference in differences; Exogeneity; Experimental design; Observational studies.

1. INTRODUCTION

There has been an outburst of work in economics that adopts the language and conceptual framework of randomized experiments. These studies, which are often called "natural experiments," examine outcome measures for observations in treatment groups and comparison groups that are not randomly assigned. Much attention is often paid to finding suitable comparison groups. This article analyzes the strengths and weaknesses of these new studies and describes how future work can be improved.

I argue that these natural experiments can be improved through the use of more complicated research designs. In particular, multiple treatment and comparison groups allow further checks of hypotheses and may allow hypotheses to be refined and alternative explanations to be ruled out. Similarly, multiple preintervention or postintervention observations can be used to examine the comparability of comparison groups and the influences of omitted factors. These and other design features can increase the validity of inferences that can be drawn from natural experiments.

Good natural experiments are studies in which there is a transparent exogenous source of variation in the explanatory variables that determine the treatment assignment. A natural experiment induced by policy changes, government randomization, or other events may allow a researcher to obtain exogenous variation in the main explanatory variables. This occurrence is especially useful in situations in which estimates are ordinarily biased because of endogenous variation due to omitted variables or selection. Such approaches have recently been used to analyze a wide range of issues. The natural-experiment approach emphasizes the general issue of understanding the sources of variation used to estimate the key parameters. In my view, this is the main lesson of these studies. If one cannot experimentally control the variation one is using, one should understand its source. This idea

is evident in past research, but natural experiments certainly give it more emphasis.

A couple of examples illustrate how a natural experiment may allow the study of the effects of exogenous variation in an explanatory variable that is in other situations endogenously related to the outcome of interest. First, in studies of the effects of social insurance programs on labor supply it is often difficult to distinguish the effects of an individual's benefit entitlement from the effects of past labor supply and earnings that typically determine that benefit entitlement. Specifically, programs such as unemployment insurance, workers' compensation, and Social Security condition eligibility and the level of benefits on previous earnings. Previous earnings are highly correlated with future earnings and the payoff to work. Thus, in studies of the effects of these programs on employment and earnings, it may be difficult to separate the independent influences of earnings history from benefit generosity. This problem is typically exacerbated by the use of proxies for the relevant earnings and benefit variables so that idiosyncratic and potentially exogenous variation in the benefit variables is often lost. Because of this concern, many recent studies have examined changes in social insurance benefits that applied to certain groups but not others. For example, Classen (1979), Solon (1985), Meyer (1989, 1992), and Green and Riddell (1993a,b) examined unemployment insurance, and Meyer, Viscusi, and Durbin (1990, in press), Krueger (1990), Gardner (1991), and Curington (1994) examined workers' compensation.

A second example is provided by studies of the effects of military service on earnings. Work that compares civilian earnings by veteran status may be biased if a nonrandom group of individuals serves in the military. In particular, those who enlist may face worse labor-market opportunities than those who do not enlist. Alternatively, military induction may screen out those individuals in worse health. Recent

work has overcome this problem by using the variation in veterans' status caused by the Vietnam-era draft lottery or the World War II draft mechanism, which depended on date of birth (Angrist 1990; Angrist and Krueger 1994). Other research uses variation across cohorts in conscription rates in the Netherlands (Imbens and van der Klaauw 1994).

Other topics that have been examined using natural experiment approaches include, though are by no means limited to, the effects of minimum wage laws through studies of changes in state laws (Card 1992a; Card and Krueger 1994) and federal law (Card 1992b; Katz and Krueger 1992); the effects of immigrants on the employment and wages of natives through studies of the impacts of large influxes of immigrants (Card 1990; Hunt 1992); the effects of family size on family choices using the delivery of twins as exogenous variation (Bronars and Grogger 1993; Rozensweig and Wolpin 1980); the effects of taxes on labor supply and investment by examining tax reforms (Cummins, Hassett, and Hubbard 1994; Eissa 1993); the effects of Medicaid on health, labor supply, and Aid to Families with Dependent Children participation through studies of program expansions that have expanded eligibility to certain groups (Currie and Gruber 1993; Yelowitz 1994) and related work on Canada (Hanratty 1992); and the effects of liquidity constraints on investment using changes in oil prices as shocks to the cash flow of nonoil subsidiaries (Lamont 1993).

There are certainly many antecedents to this literature. Examples in economics include Rose (1952), who analyzed strike rates before and after compulsory mediation laws, and Simon (1966) who examined liquor sales before and after state price increases. Both authors used comparison states that did not have law or price changes. Many of the research designs discussed here have been extensively analyzed in psychology, where they are called quasi-experiments. Because there is a rich tradition of use of these methods in psychology, I will often refer to the parallel terms and literature from that discipline. The term quasi-experiments emphasizes that such studies are not quite experiments. The term natural experiments, which is more commonly used in economics, somewhat inappropriately suggests that these studies are experiments and moreover that they are spontaneous. In economics we have not settled on a name for the approach of conventional studies, where the process that determines the variation in the key explanatory variables is not known or modeled. In psychology such studies are called correlational designs, or static-group comparisons. Such studies are based on variation that commonly occurs, usually in a cross-section.

The article proceeds as follows. Section 2 outlines some general problems, drawing inferences in empirical work. Section 3 describes the main research designs used in natural experiments. Section 4 describes extensions to these methods, and Section 5 indicates ways of probing the comparability of comparison groups. Section 6 outlines ways that the hypotheses under test can be further examined, and Section 7 describes the sources of exogenous variation in natural experiments and other studies. Section 8 describes instrumental variables methods that have been used when

treatment assignment is imprecise. Section 9 discusses how to interpret the results from natural experiments, and Section 10 concludes.

2. THREATS TO VALIDITY

It is useful to begin by outlining some of the general problems with drawing economic conclusions from empirical studies. Because these problems apply to some extent to all of the research designs discussed here, it is useful first to describe them in general. A good starting point is Donald Campbell's (Campbell 1957, 1969; Campbell and Stanley 1966; Cook and Campbell 1979) list of "threats to validity." These threats are problems that may undermine the causal interpretations in studies. The examination of threats is a study-by-study problem. A detailed knowledge of the theory, institutions, data collection, and other background relevant to a topic is necessary to judge the importance of these problems for a given study.

2.1 Threats to Internal Validity

Internal validity refers to whether one can validly draw the inference that within the context of the study the differences in the dependent variables were caused by the differences in the relevant explanatory variables. Although I have altered the list of threats and their descriptions to make them more relevant to economics, the debt to Campbell is clear:

1. Omitted variables: Events, other than the experimental treatment, occurring between preintervention and postintervention observations that provide alternative explanations for the results.

2. Trends in outcomes: Processes within the units of observations producing changes as a function of the passage of time per se, such as inflation, aging, and wage growth.

3. Misspecified variances: The overstatement of the significance of statistical tests due to effects such as the omission of group error terms that indicate that outcomes for individual units are correlated.

4. Mismeasurement: Changes in definitions or survey methods that may produce changes in the measured variables. This would include changes in survey wording or question order. For example, there have been important recent changes of this kind in the Current Population Survey (CPS) unemployment and education questions. One might also include in this category seam-bias problems in which higher levels of changes (e.g., becoming unemployed or going on welfare) are reported for periods that straddle successive interviews than are reported for analogous time periods that are surveyed in the same interview (Citro and Kalton 1993) and time-in-survey effects such as rotation-group bias in the CPS unemployment rate (Bailar 1975).

5. Political economy: Endogeneity of policy changes due to governmental responses to variables associated with past or expected future outcomes.

6. Simultaneity: Endogeneity of explanatory variables due to their joint determination with outcomes.

7. Selection: Assignment of observations to treatment groups in a manner that leads to correlation between assignment and outcomes in the absence of treatment. Selection can take many forms. For example, observations may be assigned to a treatment group based on previous extreme values of the dependent variable or variables associated with the dependent variable. In the training literature, it has been emphasized that a decline in earnings frequently precedes program entry because program operators tend to enroll those individuals with recent labor-market problems (Ashenfelter 1978; Ashenfelter and Card 1985; Heckman and Smith 1994). This rule for selecting participants makes comparisons of changes in earnings for participants and nonparticipants difficult. Different types of selection will have different remedies. For example, selection based on time-invariant individual characteristics possibly may be differenced away, but selection bias based on the lagged dependent variable could be exaggerated by this approach. The general selection problem is the subject of an extensive literature, as discussed by Heckman and Robb (1985) and Manski (1989).

8. Attrition: The differential loss of respondents from treatment and comparison groups. For example, Hausman and Wise (1979) examined attrition of individuals from a randomized experiment, and Pakes and Ericson (1990) examined attrition from a firm panel due to liquidations and changes in ownership.

9. Omitted interactions: Differential trends in treatment and control groups or omitted variables that change in different ways for treatment and control groups. An example is a time trend in a treatment group that is not present in a comparison group. The exclusion of such interactions is a common identifying assumption in the designs of natural experiments.

I should emphasize that this list is a practical list of concerns rather than an exhaustive list of possible problems. Campbell's list was modified several times in later work by Campbell and others (Cook and Campbell 1979; Cook and Shadish 1994). This later work emphasized such threats as "diffusion and imitation of treatments" and the "compensatory equalization of treatments." These ideas are similar to the observation in recent studies of training that the controls often receive some training through other programs even when they are denied entry to the program being studied.

2.2 Threats to External Validity

Cook and Campbell (1979) enumerated three threats to external validity. External validity deals with whether effects found in an experiment can be generalized to different individuals, contexts, and outcomes. In essence, these threats to external validity are just the possibility that there are important interactions between the treatment and individual characteristics, location, or time:

1. Interaction of selection and treatment: Unrepresentative responsiveness of the treated population. The treatment group may not be representative of certain population, or the treatment may be different from that which one would like to examine.

2. Interaction of setting and treatment: The effect of the treatment may differ across geographic or institutional settings.

3. Interaction of history and treatment: The effect of the treatment may differ across time periods.

Examples of such interactions come from studies in which the treatment involves changing a key explanatory variable (the workers' compensation benefit or the minimum wage) from one value to another. The effect of a given change in this explanatory variable may depend on the range of the variable over which this change is made and the composition of the treatment group. This issue is of particular concern if one seeks to extrapolate the results. This problem is not unique to natural experiments, however.

Cook and Campbell also enumerated 10 threats to what they call "construct validity." This concept refers to confusion over what is actually the cause and effect, such as confusion over the relevant part of a treatment that has many dimensions. Although these types of issues may not seem important in economics, they do arise. For example, Angrist (1990) asked to what extent the effect on earnings of being draft eligible is purely due to service in the military per se rather than the effects of draft avoidance behavior and special educational programs for veterans. A second example is the debate over signaling versus productive schooling in which it is asked whether the higher earnings of the more educated are due to the credentials that signal ability or the lessons that impart skills (Ehrenberg and Smith [1994, pp. 308–312] provided a nice summary and references).

There are other threats to the generalizability of study results that are again not unique to natural experiments. First, one might expect difficulty in extrapolating results from a temporary change to a permanent one in which individuals as well as institutions fully adapt to the new situation. Second, one might expect general equilibrium response to changes such as labor-market displacement effects. These issues and others receive much attention in the literature on social experiments (Hausman and Wise 1985, and more recently Meyer in press; Spiegelman and Woodbury 1990). See also Manski (1994) for several results on extrapolating experimental evidence from one group to another.

3. THE RESEARCH DESIGNS

Three of the main goals of a research design should be (1) finding variation in the key explanatory variables that is exogenous, (2) finding comparison groups that are comparable, and (3) probing the implications of the hypotheses under test. Without the ability to experimentally vary the relevant variables, researchers should seek to find variation that is driven by factors that are clearly identified and understood. One can then make an informed decision about the exogeneity of that variation and rule out other explanations. Being able to rule out obvious sources of endogeneity is not enough, however. The possibility of omitted variables, trends in outcomes, omitted interactions, and so forth places a burden on the researcher to examine the comparability of groups that

are being compared. Often other information from additional comparison groups or time periods can be used to examine comparability. It is also often possible to further probe hypotheses by-refining them and subjecting them to additional tests. These ideas need to be kept in mind when analyzing the design of any study. Recent work that emphasizes some of these themes includes that of Rosenbaum (1987), Cook (1991), and Cook and Shadish (1994).

There are a few study designs that have been commonly used in natural experiments. Many other works use slight variants on these designs. I begin with the simplest design.

3.1   The One Group Before and After Design

I begin with this design because it is often used as a method of preliminary analysis and because many other methods are logical extensions of this approach. In psychology, this approach has been called the one group pretest-posttest design. In economics, this approach is often called differences, based on the most common statistic calculated with such data. This approach is not very likely to lead to valid inferences, but it may be appropriate in some situations. Most of the more complicated designs are used to overcome some difficulty or deficiency with this simple design or to determine if the inferences from it are valid.

The one group before and after design is motivated by the equation

$$y_{it} = \alpha + \beta d_t + \epsilon_{it}, \qquad (1)$$

where y_{it} is the outcome of interest for unit i in period t, $t = 0, 1$, and $i = 1, \ldots, N_t$. d_t is a dummy variable for being in the treatment group—that is, $d_t = 1$ if $t = 1$ and 0 otherwise—and β is the true causal effect of the treatment on the outcome for this group. The treatment group is usually defined (at least in part) by the variation in another variable such as the level of the minimum wage or the workers' compensation benefit. Examples of outcomes include employment in the minimum-wage studies or time out of work in the workers' compensation studies.

The key identifying assumption of this model is that, in the absence of the treatment, β would be 0; that is, there would be no difference in the mean of those in group 0 and group 1. This condition is typically written as $E[\epsilon_{it} \mid d_t] = 0$; that is, the conditional mean of the error term does not depend on the value of the treatment dummy. Using a term common in statistics, one might say that this condition is implied by ignorable treatment assignment (Rubin 1978). If this condition holds, an unbiased estimate of β can be obtained as

$$\widehat{\beta}_d = \Delta \bar{y}$$
$$= \bar{y}_1 - \bar{y}_0, \qquad (2)$$

where the bar indicates an average over the individual units and the subscript on y denotes the time period. Under typical assumptions, $\widehat{\beta}_d$ would also be consistent as the number of units in each group goes to infinity. β can also be obtained by directly estimating the parameters of Equation (1) using pooled data from the two time periods. This regression approach will reproduce β_d and, if one allows the variance of ϵ_{it} to vary with t, give the same standard error.

Although I focus on analyzing the mean difference between the treatment group and the comparison group, other summary measures of the differences in the distributions may be of interest. These measures include quantiles such as the median or 75th percentile and the cumulative distribution function at certain points. For example, Meyer et al. (in press) examined both the cumulative distribution function and quantiles of the outcome variable, injury duration. When examining the effect of an explanatory variable on years of education received, we may be interested in the fraction of a group graduating from high school (a point on the cumulative distribution function).

Because nothing so far requires data on individual units rather than grouped means, one could use aggregate data for this pretest-posttest approach. One would need an estimate of the variances of any statistics examined, however, to conduct statistical tests. Some other advantages of individual-level data are discussed later. If individual data are used, the samples could be different in the two periods; that is, one might use repeated cross-section data rather than panel data.

The use of the one group before and after design requires very special circumstances. One needs strong evidence that the two groups would have been comparable over time in the absence of the treatment. An example that illustrates this issue is the work of Meyer et al. (1990, in press). These authors examined the effect of the level of workers' compensation benefits on the length of time out of work by comparing individuals injured during the year before and after two large increases in state maximum weekly benefit amounts. These increases raised the benefit amount for an easily identified class of high-income workers who were injured after the changes in the state laws. In this example, there might be other influences on injury duration that one would want to rule out as explanations for any outcomes. One influence would be other changes in the law regarding injury compensation over the study period (omitted variables). Similarly, one might be concerned about overall trends in injury severity or changes in the reporting practices of the insurance companies that submit the records (trends in outcomes or mismeasurement). If the data were from a panel, one might be worried about nonrandom exit from the sample (attrition). One way to assess the importance of these threats to internal validity is to examine the outcomes for similar groups that did not receive the treatment but would presumably be subject to these influences as well. Such an idea leads to the next design.

3.2   The Before and After Design With an Untreated Comparison Group

Often data will be available for the time period before and after the treatment for a group that does not receive the treatment but experiences some or all of the other influences that affect the treatment group. When such a group is present, the design in psychology has been called the untreated control group design with pretest and posttest. In economics the approach is identified with the most common statistical technique used in this situation, difference in differences.

When one has a comparison group over the same time period as the before and after groups, often the underlying model of the outcome variable is of the form.

$$y_{it}^j = \alpha + \alpha_1 d_t + \alpha^1 d^j + \beta d_t^j + \epsilon_{it}^j, \qquad (3)$$

where the outcome y is now also indexed by j for the group, $j = 0, 1$, and $d_t = 1$ if $t = 1$ and 0 otherwise, $d^j = 1$ if $j = 1$ and 0 otherwise, and $d_t^j = 1$ if $t = 1$ and $j = 1$ and 0 otherwise. d_t^j is a dummy variable for being in the experimental group after it receives the treatment, and β is the true causal effect of the treatment on the outcome for this group. Again, the key identifying assumption is that β would be 0 in the absence of the treatment, or $E[\epsilon_{it}^j \mid d_t^j] = 0$. In this case, and unbiased estimate of β can be obtained by difference in differences as

$$\begin{aligned} \widehat{\beta}_{dd} &= \Delta \bar{y}_0^1 - \Delta \bar{y}_0^0 \\ &= \bar{y}_1^1 - \bar{y}_0^1 - (\bar{y}_1^0 - \bar{y}_0^0), \end{aligned} \qquad (4)$$

where again a bar indicates an average over i, the subscript denotes the time period, and the superscript denotes the group. The key idea behind this approach is that α_1 summarizes the way that both group $j = 0$ and group $j = 1$ are influenced by time. There may be a time-invariant difference in overall means between the groups $j = 0$ and $j = 1$, but this aspect is captured by α^1.

This research design is the essence of two recent studies. Card and Krueger (1994) examined the effects of an increase in the New Jersey state minimum wage on employment. Their sample consists of fast-food restaurants from four chains in New Jersey before ($t = 0$) and after ($t = 1$) the increase in the minimum wage. In addition, they examined employment at a sample of similar restaurants in eastern Pennsylvania over the same time period. This sample from Pennsylvania provides a group ($j = 0$) that is plausibly subject to the same changes over time as the group in New Jersey, except that Pennsylvania did not change the minimum wage. These common changes are captured by α_1 in Equation (3). This term represents such things as macroeconomic conditions and regional growth trends in fast-food employment over the period.

Meyer et al. (1990, in press) examined the effects of two large workers' compensation benefit increases on the length of claims. They also relied on an untreated comparison group, as well as before and after groups. The untreated comparison group is those individuals within a state who were not subject to the increases in workers' compensation benefits because they had average or low earnings. These comparison workers were likely to be subject to any other changes in program administration or insurers' claim-monitoring procedures.

Again, β can be estimated directly by applying ordinary least squares to Equation (3). This method reproduces $\widehat{\beta}_{dd}$ but gives a different standard error for the estimate unless one allows the error variance to differ across the four groups defined by t and j. An advantage of the regression formulation is that it makes clear that the key identifying assumption is that there is no interaction between $t = 1$ and $j = 1$ (except for the influence under study).

Several of the internal validity threats are reduced by this approach, but important concerns may remain. In the workers' compensation and minimum wage studies such influences as changes in other state laws and labor-market conditions (omitted variables) and any changes in surveyors' methods (mismeasurement) are likely reduced by the use of the untreated comparison group. The importance of trends in employment and the duration of workers' compensation receipt (trends in outcomes) is also reduced or eliminated. The comparability of the before and after groups is higher if sample attrition is negligible. Card and Krueger (1994) went to great lengths to determine the status and employment of nonresponding establishments in their panel study. In repeated cross-section studies one usually does not have an attrition problem, but one needs to examine if the samples are selected over time in the same way from comparable populations.

One of the main threats to the validity of inferences from this design is the possibility of an interaction (besides the treatment) between $j = 1$ and $t = 1$ (omitted interactions). Changes in other state laws or macroeconomic conditions are not likely to always influence all groups in the same way. A recession may have a disproportionate effect on one income group compared to another or in one state than another. This design is most plausible when the untreated comparison group is very similar to the treatment group so that interactions are less likely.

A situation favorable to this design is one in which the comparison group both before and after has a distribution of outcomes close to that for the treatment group during the before period. If there are large differences, then transformations of the dependent variable in (3) may affect the results. For example, if the mean of the outcome variable is very different in the treatment and comparison groups, then (3) could not be an appropriate model both in levels and logarithms (unless $\alpha_1 = 0$). This problem occurs because nonlinear transformations of the dependent variable imply different marginal effects on the dependent variable at different levels of the dependent variable. Thus time could not have an effect of the exact same magnitude in both treatment and control groups in both a linear and logarithmic specification. In this example, one may be able to determine whether a linear or logarithmic transformation is more appropriate by testing whether other variables change the dependent variable in a linear or a logarithmic fashion. The same issue arises when the right side of (3) is nonlinear or when maximum likelihood techniques are used (e.g., see Madrian 1994). In any case, it is useful to examine the size and significance of $\widehat{\alpha}_1$ and $\widehat{\alpha}^1$ for an indication of the comparability of the groups. If they are both near 0, then possible transformations of the dependent variable or different functional forms in likelihoods should be of little importance.

In addition, examining the size and significance of $\widehat{\alpha}_1$ and $\widehat{\alpha}^1$ may reveal other problems of interpretation. If $\widehat{\alpha}_1$ is large in absolute value, it suggests that period-to-period changes in the dependent variable are not unusual and further evidence on its variance over time might be warranted. If the effects of omitted variables, trends in outcomes, mismeasurement,

and so forth that are captured by this term are large, it is more likely that the effect varies substantially across groups. A large $\widehat{\alpha}_1$ may also be an indication that standard errors are understated due to the presence of a group effect in the error term for the interaction of treatment and time.

Although these last remarks are given as a rule of thumb rather than an absolute principle, the idea can be formalized in the following way. Suppose that we are willing to assume that α_1 represents the average effect of a change in an unobserved explanatory variable that has a heterogeneous but positive effect on the outcome for all observations. Then a bound on the true interaction between time and being after the treatment is α_1/p, where p is the fraction of the after population that is in the treatment group. Because this bound is decreasing in the size of α_1, we are able to rule out interactions of a smaller magnitude the smaller α_1 is. A similar argument can be made about α^1.

I should note that the appropriate error structure in (3) may differ in repeated cross-section and panel data. In repeated cross-section data it is likely that ϵ_{i0} is uncorrelated with ϵ_{i1} so that the averages in (4) are independent. In panel data, correlation is likely, but an easy solution is to estimate (3) in differences and in (4) calculate the variance of $\widehat{\beta}_{dd}$ using the sample variance of the quantity $(y_{i0} - y_{i1})$. In the case of positively correlated ϵ's the variance of the difference is smaller than the sum of the individual variances.

4. EXTENSIONS OF DIFFERENCE–IN–DIFFERENCES METHODS

To narrow the focus of this article, I omit some research designs that have found use in other fields, such as the regression discontinuity design (Cook and Campbell 1979). The main ideas of Section 3, however, are imbedded in other commonly used research designs. This section describes several extensions to the difference-in-differences approach.

4.1 Studies Without a Time Dimension

There are many ways that the variables in Equation (3) can be relabeled without changing the underlying approach. The index t does not need to indicate time. Rather, it only needs to indicate one group that was subject to a treatment and another group that was not. For example, Madrian (1994) examined the effects of insurance coverage on the probability of moving between jobs. The hypothesis is that those with *both* current coverage and a greater demand for insurance (due to lack of coverage through a spouse or greater demand for health care due to pregnancy or large family size) should be less likely to move. Let $t = 0$ for someone with a low demand for insurance, and $t = 1$ for someone with a high demand. Similarly, let $j = 0$ for an uncovered worker, and $j = 1$ for a person currently covered. The treatment effect is the interaction of being currently covered and having a greater demand for future insurance $(t = 1, j = 1)$.

A word of caution is appropriate here. When t does not indicate being before or after an event (often a sudden change in an explanatory variable), it may be more difficult to assess whether there would be an interaction between $t = 1$ and

$j = 1$ even without the treatment. One would like to be able to examine if the outcome measure for the treatment and control groups would change by the same amount in response to differences analogous to those that define the treatment but in the case in which the treatment is not present. In the preceding example, one needs to consider if greater insurance demand as reflected in no spousal coverage, pregnancy, or large family size would have the same quantitative effect on the mobility of those with and without their own coverage even if health insurance were not an influence. In this situation it may also be more difficult or impossible to find additional observations on analogous units (time periods, states) to examine if in other contexts the mobility of those with and without coverage moves in parallel. When the units are time or states, one may be able to select similar states or additional time periods to examine this hypothesis.

4.2 Controls for Individual Characteristics

The incorporation of the influences of other variables is straightforward in the regression approach of Equation (3). If we have a vector of characteristics of the units under study, z_{it}^j, we can include it as an additional vector of explanatory variables. Thus the regression equation

$$y_{it}^j = \alpha + \alpha_1 d_t + \alpha^1 d^j + \beta d_t^j + z_{it'}^j \delta + \epsilon_{it}^j \qquad (5)$$

provides a simple way to adjust for observable differences between the observations in the different groups. Using this equation may also improve the efficiency of the estimate of β by reducing the residual variance. I should note that, as usual, enforcing homoscedasticity of the error term across groups (even if it truly holds) does not improve asymptotic efficiency.

I should also note that one needs to enforce equality of δ across groups; otherwise Equation (5) will not adjust for differences in these variables across groups. One can test whether this restriction holds using conventional methods. If the variables have different effects within the different groups, it is unlikely that the regression adjustment will eliminate these differences. A test of equality of δ across groups might also detect omitted variables or functional-form misspecifications that would make the regression adjustments inadequate.

4.3 Treatments That Are Higher-Order Interactions

In the examples so far, the treatment group has been defined by the interaction of two dummy variables, usually a dummy variable for being in the treatment group and one for being after the time of the treatment. Situations often arise in which the treatment is defined by the interaction of more than two variables. In this case, a design relying on this higher level of interaction may allow the researcher to remove main effects and lower-level interactions effects. More concretely, the researcher may believe that there are extra terms in (3) besides time and state in the United States, for example. It may be that the treatment group affects a certain demographic group in the state and time period. Thus a version of (3) may be appropriate with a higher-order interaction being the key explanatory variable with the coefficient β. This approach is suitable if the treatment group differs from the comparison group along

several dimensions, and it may have the advantage of removing any trends along these other dimensions of the data.

The regression equation for this model is

$$y_{it}^{jk} = \alpha + \alpha_1 d_t + \alpha^1 d^j + \gamma^1 e^k + \alpha_1^1 d_t^j$$
$$+ \gamma_1^1 e_t^k + \alpha^{11} d^{jk} + \beta d_t^{jk} + \epsilon_{it}^{jk}, \quad (6)$$

where the outcome y is now also indexed by k, $k = 0, 1$, and $d_t = 1$ if $t = 1$ and 0 otherwise; $d^j = 1$ if $j = 1$ and 0 otherwise; $e^k = 1$ if $k = 1$ and 0 otherwise; d_t^j, e_t^k, and d^{jk} are the three possible interactions of two factors (the first-order interactions); and $d_t^{jk} = 1$ if $t = 1, j = 1$, and $k = 1$ and 0 otherwise is the interaction of all three factors (the second-order interaction). d_t^{jk} is a dummy variable for being in the subset of the experimental group that receives the treatment after it receives the treatment, and β is the effect of the treatment on the outcome.

Examples of such designs include that of Gruber (1994), who examined the incidence of mandated maternity benefits, and Yelowitz (1994) who examined the effects of Medicaid expansions on welfare participation and labor supply. In the work of Gruber (1994), the treated are those women of certain ages ($k = 1$) in a certain group of states ($j = 1$) after the mandate ($t = 1$). The coefficient on this second-order interaction is the key parameter of interest. Variables to capture the main effects and first-order interactions are also included in the estimation equations. Similarly, in the work of Yelowitz (1994) the treated are mothers with children of certain ages ($k = 1$) in certain states ($j = 1$) after extensions of Medicaid coverage ($t = 1$). Again, the coefficient on a second-order interaction variable is the key parameter of interest. This idea can be extended to even higher-level interactions. It is important to include the first-order interactions in Equation (6) when testing for the presence of the second-order interaction; otherwise the second-order interaction effect would be confounded with the omitted first-order interactions, likely leading to biased estimates.

5. FURTHER EVIDENCE ON COMPARABILITY

The use of the before and after design with an untreated comparison group rests on comparability of the before and after groups, at least after netting out a time mean common to both the treatment and comparison groups. To examine comparability, often supplementary information is available. Examples of supplementary information include a clearly specified hypothesis about the likely differences between the before and after groups or additional control groups.

5.1 Multiple Comparison Groups

The before and after design with an untreated comparison group can be strengthened by the use of additional comparison groups. This design feature allows further examination of the $\beta = 0$ hypothesis in the absence of a treatment. Additional comparison groups reduce the importance of biases or random variation in a single comparison group. There are some simple principles to follow in choosing comparison groups. The more similar the comparison group is to the treatment group the better. For a given degree of similarity

with the treatment group, however, greater differences across comparison groups are desirable if they are likely to lead to different biases.

Last, the more comparison groups the better. Examples of studies that feature multiple comparison groups include that of Meyer (1989), who examined the effect of 17 increases in unemployment insurance benefits on unemployment durations and 16 analogous cases in which benefits were unchanged. Similarly, Krueger (1990) examined the effect of an increase in workers' compensation benefits on claim durations and used two groups of comparison workers who were not subject to the benefit increase.

Another approach may be available if the researcher has knowledge about how the treatment and comparison groups differ, say that one group has a higher mean value of a given, possibly unmeasured, variable that affects the outcome. One can examine if groups that differ in the mean value of this variable respond to other factors (time, for example, in the before and after designs) similarly. This idea has been called "control by systematic variation" (see Rosenbaum 1987 for a nice discussion). If the groups do respond similarly, it would support the assumption of no omitted interactions and the converse if they do not.

When a comparison group that would be expected to be similarly affected by other factors cannot be identified, one possible approach is to search for comparison groups whose outcomes could be expected to bracket the outcome for the treatment group. For example, using groups that might be expected to have both larger and smaller responses to other changes during the relevant time period could provide bounds for the possible effects. If these bounds are narrow, the method provides useful information about the parameter of interest.

In related work, Rosenbaum (1987) emphasized that one can formally test whether comparison groups are similar to each other. This approach is likely to be most useful when the assumptions that make one group a valid comparison group imply that the other is valid also. Then a comparison of the two groups provides a test of these assumptions. This comparison of control groups can be reinterpreted as the economist's test of overidentifying restrictions. A comparison of two estimates of β, based on a summary statistic from the same treatment group but different comparison groups, would just be a test of equality of the summary statistic from the two comparison groups.

Additional comparison groups are sometimes called null treatment groups. A null treatment group is a group like the treatment group in time or geography that does not receive a treatment. The use of such groups to check assumptions regarding the unbiasedness or variance of outcome measures has been called "uniformity trials" (see Margolin 1987).

5.2 Multiple Preintervention or Postintervention Time Periods

A design feature that allows the examination of various validity threats is the use of data from several preintervention or postintervention time periods. For example, Meyer

(1992) studied the effect of a 27% increase in the New York unemployment insurance benefit that only applied to above-average-income workers. An analysis of two years of quarterly data makes it clear that there is strong seasonality (omitted variables in terms of the validity threats) in the outcome measures, which are the number and length of claims. This result suggests that comparing outcomes for the months immediately before and after the benefit increase would be inappropriate. Additional time series observations may also indicate the importance of group effects in the error of equations such as (5) (misspecified variances).

A second and underemphasized advantage of a long time series for outcome measures is that they may allow the researcher to examine if the treatment and control groups tend to move in parallel—that is, go up or down together. In the absence of interactions between treatment and other influences (omitted interactions), parallel movement would be expected. In the case of the minimum wage study described previously, one could ask if the New Jersey and Pennsylvania fast-food employment levels tend to move together. One could also examine if movements of a given magnitude are more or less common than the standard errors suggest.

5.3 Other Ways to Examine Comparability

In the before and after comparison designs, with or without an untreated group, it is useful to compare the characteristics of the units in the groups for indications of comparability. If substantial differences in mean characteristics are present, they should cause concern. Such differences would not necessarily invalidate the results because one may be able to control for these characteristics. In addition, group differences in characteristics may accord with or reject hypotheses about likely differences between treatment and comparison groups. Similarly, in repeated cross-section analyses the population (or sample) sizes should be examined because large changes in the size of the population may indicate changes in the composition of the groups in ways not completely captured by observed characteristics. The direction of change, especially if competing hypotheses imply directions, may be informative.

Last if assignment to the treatment and comparison groups is defined by a measured variable, it may be possible to examine comparability in an additional way. In the preceding unemployment insurance and workers' compensation examples, the groups were defined by past earnings. One could look for an effect of earnings on the outcome of interest *within* the high and low earnings groups. If there is no relationship, that suggests that the groups are more likely to be comparable, and the opposite is true if there is a strong relationship.

6. FURTHER TESTS OF THE HYPOTHESES

6.1 Multiple Treatment Groups

The before and after design with an untreated comparison group is also strengthened by the presence of several distinct groups that are subject to the treatment. Especially useful are treatment groups in different settings such as different time periods or states or treatment groups receiving treatments of different intensities. This design feature was used by Meyer et al. (1990, in press), who examined the same sets of statistics for workers' compensation benefit increases in two different states. Differences in the intensity of the treatment across different groups allow one to examine if the changes in outcomes differ across treatment levels in the expected direction. Multiple treatment groups may also allow testing of more refined hypotheses if the treatment is expected to have a differential impact on the outcomes for different groups. For example, Card and Krueger (1994) in their minimum wage study defined additional treatment groups within New Jersey defined by high, medium, and low wages prior to the increase in the minimum. The expectation is that the minimum wage would have a larger effect in restaurants with lower wages.

6.2 Other Tests

Another design feature that strengthens evidence of causal effects is the later reversal of the initial treatment. For example, a state law may be passed and then later repealed. Curington (1994) examined several changes in workers' compensation benefits that increased the worker's reward for lengthening the period of receipt of temporary benefits relative to permanent benefits. He was also able to examine one change that reversed the incentives, increasing permanent benefits relative to temporary ones, and saw if the effect on the dependent variable reversed. In a slightly different vein, Gruber (1994) examined a later federal mandate of maternity benefits that changed some of his earlier treatment states (which had state mandates) into controls, and vice versa.

A final way of examining the appropriateness of an approach (especially when no effects are found) is to examine if the approach has sufficient power to detect and properly measure the effect of a known causal variable on the outcome. Margolin (1987) called this the use of a positive treatment group.

7. SOURCES OF EXOGENOUS VARIATION

One of the themes from the examples that I have described is that government policies often create natural treatment and comparison groups. Frequently, this event occurs because our federal system of government allows one state to change a policy while others do not. The many cross-state differences in policies and changes in these policies allow the examination of a wide range of questions.

An illustrative example comes from studies of the effects of unemployment insurance and workers' compensation benefits on the length of absence from work. These studies have typically used several sources of variation in benefits generally in unspecified proportions. These sources include (a) the variation due to differences in individuals' earnings histories that is often the sole determinant of benefits within a state at a point in time, (b) differences across states in these schedules that relate current benefits to past earnings, and (c) changes in these benefit schedules over time.

Given that labor supply is correlated over time for an individual, (a) is unlikely to be exogenous. In terms of the validity threats, there is selection on a variable highly correlated with

the lagged dependent variable. Because it is difficult to measure characteristics of state programs and state labor markets that are important for the outcomes of interest, mean differences across states incorporated in (b) are unlikely to be useful. Again, in terms of the validity threats, there are likely important omitted variables. This leaves interactions between (a) and (b)—that is, how different states treat differences in earnings history differently—and (c) as possible sources of variation. Interactions between (a) and (b) in addition to (c) are the sources of variation used by Meyer (1990) and Anderson and Meyer (1994). Source (c) is likely to be especially useful if one can examine sharp changes in policies that were unlikely to have been determined by past values of the outcomes of interest (so that political economy issues are not important). It is also important that past values of outcomes could not have been affected by knowledge of future policy changes. Conventional studies typically include all three sources of variation in unknown proportions and do not adequately control for the sources of endogeneity (i.e., the selection mechanism or omitted variables) of the different components. Recent natural experiment studies have taken the more convincing approach of examining changes in benefits, typically ones that affect some groups but not others.

Nevertheless, not every law change is a good natural experiment. The danger in using such variation is that the changes may be driven by political factors associated with outcomes. Campbell (1969) and Cook and Tauchen (1982) provided a good summary of the argument in the case of natural experiment approaches, and Besley and Case (1994) provided a recent discussion. For example, a few high years of crime due to unusual circumstances may stimulate a crackdown. A subsequent reduction in crime after the unusual years should not be taken to indicate an effective crackdown if a drop would have been expected anyway.

The way to avoid that pitfall is to know the circumstances surrounding reforms or to more generally model the determinants of policy changes. For example, Cook and Tauchen (1982) examined whether liquor taxes affect the incidence of heavy drinking. Because of a concern that changes in drinking habits might cause changes in taxes, they performed a series of Granger and Sims exogeneity tests to see if there is evidence of a causal link running in this other direction.

Other situations favorable to the use of natural experiment approaches include changes in government policies that are applicable to some groups but not others. Then, unless past changes specific to that group motivated the policy change, the groups not subject to the policy change provide comparison groups. Additionally, data from before and after large, sharp policy changes can often avoid the influence of slowly moving factors that determine political decisions.

8. IMPRECISE ASSIGNMENT TO TREATMENT AND INSTRUMENTAL VARIABLES

Government policy or other forces do not always create simple treatment and comparison groups but may instead influence the likelihood that an individual receives a treatment. In such a situation, a variable that is correlated with the as-

signment to treatment but that does not solely determine the assignment may be available. If this variable is also uncorrelated with the error in the outcome equation, then in linear and certain nonlinear models the effects of the treatment may be estimated using instrumental variables (IV).

An example of the imprecise assignment to treatment and the use of IV comes from the Vietnam-era draft lottery. The lottery randomly established priority for induction into the military. Draft lottery numbers can be used as instruments for veteran status when one is interested in determining the effect of military service on earnings later in life (Angrist 1990). A second example is the work of Angrist and Krueger (1991), who used quarter of birth as an instrument for educational attainment in earnings equations. Quarter of birth is correlated with educational attainment because it affects whether compulsory school attendance laws were binding. The authors also provided substantial evidence that quarter of birth does not have an effect on earnings except through compulsory schooling laws.

IV estimation also may often be applied to a modified version of the earlier equations. Changes in state laws can be thought of as generating instruments that can be used to identify causal effects. For example, in the workers' compensation study described at length previously, one could include as an explanatory variable the benefit amount, b_{it}^j, and then use the appropriate dummy variable as an instrument. The equation would become

$$y_{it}^j = \alpha + \alpha_1 d_t + \alpha^1 d^j + \pi b_{it}^j + z_{it'}^j \delta + \epsilon_{it}^j, \qquad (7)$$

and d_t^j would be the appropriate instrument for the benefit amount. The first-stage regression in two-stage least squares (2SLS) consists of regressing the benefit amount on the dummy variable d_t^j and other control variables. The predicted values from this regression are then substituted for the benefit amount in the second stage of 2SLS. An advantage of the IV approach in this example is that one directly estimates the derivative of y with respect to the benefit amount.

9. WHAT PARAMETER DO NATURAL EXPERIMENTS ESTIMATE?

A weakness of natural experiments is that their results may not be generalizable beyond the group of individuals or firms or the setting used in the study. The workers' compensation example described earlier estimated the effect of increasing workers' compensation benefits over a particular range for a group of high-wage workers. One might wonder if these results generalize to all workers. Similarly, the minimum wage study described earlier estimated the effect of minimum wage laws on employment for a group of four national chains of fast-food restaurants. One might wonder if these results generalize to all low-wage employment.

This issue of external validity arises when the treatment has different effects on different observations. In terms of the equations, β in (5) may vary over i and may interact with z_i. Interactions between the treatment and individual characteristics have been examined in some natural experiments. For example, Angrist (1990) and Angrist and Krueger

(1991) performed most of their analyses separately for blacks and whites. In most cases, however, researchers permit little or no variation in treatment effects.

Although this problem certainly limits the generalizability of natural experiment results, many (or most) conventional studies assume constant impacts across groups or constant elasticities. Furthermore, in conventional studies one rarely determines what parts of the variation in a key explanatory variable are particularly influential—that is, play the largest roles in determining coefficient values. This variation is rarely the unweighted variation in a sample because a multiple regression coefficient is determined by the residual variation after other explanatory variables are conditioned out. In conventional studies, typically all sources of variation are combined and treated equally, even though some of the sources may be endogenous or have different effects on the outcomes. In natural experiments one usually can point to the source of the variation that generated the results.

10. CONCLUSIONS

Policy change, government randomization, or other events may allow a researcher to obtain exogenous variation in key explanatory variables. This is especially useful when there is concern about omitted variables, purposeful selection into treatment, or other threats to validity. Research designs based on exogenous variation have recently been used to analyze a wide range of issues. Even when not conclusive, the simplicity of such designs often narrows the range of plausible alternative explanations.

Of course, calling a source of variation a natural experiment does not make that variation exogenous. But the natural experiment approach emphasizes the importance of understanding the source of variation used to estimate key parameters. In my view, this is the primary lesson of recent work in the natural experiment mold. If one cannot experimentally control the variation one is using, one should understand its source.

ACKNOWLEDGMENTS

REFERENCES

Anderson, P. M., and Meyer, B. D. (1994), "Unemployment Insurance Benefits and Takeup Rates," Working Paper 4787, National Bureau of Economic Research, Cambridge, MA.

Angrist, J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records," American Economic Review, 80, 313–336.

Angrist, J. D., and Krueger, A. B. (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?" Quarterly Journal of Economics, 106, 979–1014.

——— (1994), "Why Do World War II Veterans Earn More Than Nonveterans?" Journals of Labor Economics, 12, 74–97.

Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings," Review of Economics and Statistics, 60, 47–57.

Ashenfelter, O., and Card, D. (1985). "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," The Review of Economics and Statistics, 67, 648–660.

Bailar, B. A. (1975), "The Effects of Rotation Group Bias on Estimates From Panel Surveys," Journal of the American Statistical Association, 70, 23–30.

Besley, T., and Case, A. (1994), "Unnatural Experiments? Estimating the Incidence of Endogenous Policies," working paper, Princeton University, Dept. of Economics.

Bronars, S., and Grogger, J. (1993), "The Socioeconomic Consequences of Teenage Childbearing: Findings From a Natural Experiment," Family Planning Perspectives, 25, 156–162.

Campbell, D. T. (1957), "Factors Relevant to the Validity of Experiments in Social Settings," Psychological Bulletin, 54, 297–312.

——— (1969), "Reforms as Experiments," American Psychologist, 24, 409–429.

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

Card, D. (1990), "The Impact of the Mariel Boatlift on the Miami Labor Market," Industrial and Labor Relations Review, 43, 245–257.

——— (1992a), "Do Minimum Wages Reduce Employment? A Case Study of California 1978–89," Industrial and Labor Relations Review, 46, 38–54.

——— (1992b), "Using Regional Variation in Wages to Measure the Effects of the Federal Minimum Wage," Industrial and Labor Relations Review, 46, 22–37.

Card, D., and Krueger, A. B. (1994), "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania," American Economic Review, 84, 772–793.

Citro, C. F., and Kalton, G. (1993), The Future of the Survey of Income and Program Participation, Washington, DC: National Academy Press.

Classen, K. P. (1979), "Unemployment Insurance and Job Search," in Studies in the Economics of Search. eds. S. A. Lippman and J. J. McCall, Amsterdam: North-Holland, pp. 191–219.

Cook, P. J., and Tauchen, G. (1982), "The Effect of Liquor Taxes on Heavy Drinking," The Bell Journal of Economics, 13, 379–390.

Cook, T. D. (1991), "Clarifying the Warrant for Generalized Causal Inferences in Quasi-Experimentation," in Evaluation and Education at Quarter Century, ed. M. S., McLaughlin and D. Phillips, Chicago: University of Chicago Press, pp. 115–144.

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

Cook, T. D., and Shadish, W. R. (1994), "Social Experiments: Some Developments Over the Past Fifteen Years," Annual Review of Psychology, 45, 545–580.

Cummins, J. G., Hassett, K. A., and Hubbard, R. G. (1994), "A Reconsideration of Investment Behavior Using Tax Reforms as Natural Experiments," working paper, Columbia University, Dept. of Economics.

Curington, W. P. (1994), "Compensation for Permanent Impairment and the Duration of Work Absence: Evidence From Four Natural Experiments," Journal of Human Resources, 29, 888–910.

Currie, J., and Gruber, J. (1993), "Saving Babies: The Efficacy and Cost of Recent Expansion of Medicaid Eligibility for Pregnant Women," working paper, Massachusetts Institute of Technology, Dept. of Economics.

Ehrenberg, R. G., and Smith, R. S. (1994), Modern Labor Economics, New York: HarperCollins.

Eissa, N. (1993), "Taxation and Labor Supply of Married Women: The Tax Reform Act of 1986 As a Natural Experiment," Working Paper, Harvard University, Dept. of Economics.

Gardner, J. A. (1991). Benefit Increases and System Utilization: The Connecticut Experience, Cambridge, MA: Workers Compensation Research Institute.

Green, D. A., and Riddell, W. C. (1993a), "Qualifying for Unemployment Insurance: An Empirical Analysis," working paper, University of British Columbia, Dept. of Economics.

———— (1993b), "The Economics of Unemployment Insurance in Canada: An Empirical Analysis of UI Disentitlement," *Journal of Labor Economics*, 11, S96–S147.

Gruber, J. (1994), "The Incidence of Mandated Maternity Benefits," *American Economic Review*, 84, 622–641.

Hanratty, M. J. (1992), "Canadian National Health Insurance and Infant Health," working paper, Cornell University, Dept. of Economics.

Hausman, J. A., and Wise, D. A. (1979), "Attrition Bias in Experimental and Panel Data: The Gary Negative Income Maintenance Experiment," *Econometrica*, 47, 445–473.

———— (1985), *Social Experimentation*, Chicago: University of Chicago Press.

Heckman, J. J., and Robb, R., Jr. (1985), "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, eds. J. J. Heckman and B. Singer, Cambridge, U.K.: Cambridge University Press, pp. 156–246.

Heckman, J. J, and Smith, J. A. (1994), "Ashenfelter's Dip and the Determinants of Participation in a Social Program: Separating the Impact of Administrative Rules and Individuals Behavior," working paper, University of Chicago, Dept. of Economics.

Hunt, J. (1992), "The Impact of the 1962 Repatriates from Algeria on the French Labor Market," *Industrial and Labor Relations Review*, 45, 556–572.

Imbens, G., and van der Klaauw, W. (1994)., "Evaluating the Cost of Conscription in the Netherlands," working paper, Harvard University, Dept. of Economics.

Katz, L. F., and Krueger, A. B. (1992), "The Effect of the Minimum Wage on the Fast Food Industry," *Industrial and Labor Relations Review*, 46, 6–21.

Krueger, A. B. (1990), "Workers' Compensation Insurance and the Duration of Workplace Injuries," Working Paper 3253, National Bureau of Economic Research, Cambridge, MA.

Lamont, O. (1993), "Cash Flow and Investment: Evidence From Internal Capital Markets," working paper, Massachusetts Institute of Technology, Dept. of Economics.

Madrian, B. C. (1994), "Employment-Based Health Insurance and Job Mobility: Is There Evidence of Job-Lock?" *Quarterly Journal of Economics*, 109, 26–54.

Manski, C. F. (1989), "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 341–360.

———— (1994), "Learning About Social Programs From Experiments With Random Assignment of Treatments," working paper, University of

Wisconsin-Madison, Dept. of Economics.

Margolin, B. H. (1987), "Comment: The Use of Multiple Control Groups in Designed Experiments," *Statistical Science*, 2, 308–310.

Meyer, B. D. (1989), "A Quasi-Experimental Approach to the Effects of Unemployment Insurance," Working Paper 3159, National Bureau of Economic Research, Cambridge, MA.

———— (1990), "Unemployment Insurance and Unemployment Spells," *Econometrica*, 58, 757–782.

———— (1992), "Quasi-Experimental Evidence on the Effects of Unemployment Insurance from New York State," working paper, Northwestern University, Dept. of Economics.

———— (in press), "Lessons From the U. S. Unemployment Insurance Experiments," *Journal of Economic Literature*, 33.

Meyer, B. D., Viscusi, W. K., and Durbin, D. L. (1990), "Workers' Compensation and Injury Duration: Evidence From a Natural Experiment," Working Paper 3494, National Bureau of Economic Research, Cambridge, MA.

———— (in press), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review*, 85.

Pakes, A., and Ericson, R. (1990), "Empirical Implications of Alternative Models of Firm Dynamics," working paper, Yale University, Dept. of Economics.

Rose, A. M. (1952), "Needed Research on the Mediation of Labor Disputes," *Personal Psychology*, 5, 187–200.

Rosenbaum, P. R. (1987), "The Role of a Second Control Group in an Observational Study" (with comments), *Statistical Science*, 2, 292–316.

Rosensweig, M., and Wolpin, K. (1980), "Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment," *Econometrica*, 48, 227–240.

Rubin, D. B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

Simon, J. (1966), "The Price Elasticity of Liquor in the U. S. and a Simple Method of Determination," *Econometrica*, 34, 193–205.

Solon, G. (1985), "Work Incentive Effects of Taxing Unemployment Benefits," *Econometrica*, 53, 295–306.

Spiegelman, R. G., and Woodbury, S. A. (1990), "Controlled Experiments and the Unemployment Insurance System," in *Unemployment Insurance: The Second Half-Century*, eds. W., L. Hansen and J. F. Byers, Madison: University of Wisconsin Press, pp. 355–394.

Yelowitz, A. S. (1994), "The Medicaid Notch, Labor Supply and Welfare Participation: Evidence from Eligibility Expansions," working paper, Massachusetts Institute of Technology, Dept. of Economics.