

**Assessing Bias in the Estimation of Causal Effects: Rosenbaum  
Bounds on Matching Estimators and Instrumental Variables  
Estimation with Imperfect Instruments**

Thomas A. DiPrete  
Department of Sociology, Duke University  
Box 90088, Durham, NC 27708-0088  
Tel. (919) 660-5612 / 5614; Fax. (919) 660-5623  
Email [tdiprete@soc.duke.edu](mailto:tdiprete@soc.duke.edu)

and

Markus Gangl  
Social Science Centre Berlin (WZB)  
Reichpietschufer 50, D-10785 Berlin, Germany  
Tel. +49.30.25491.141; Fax +49.30.25491.222  
Email [gangl@wz-berlin.de](mailto:gangl@wz-berlin.de)

January 8, 2004

We acknowledge support provided by the Wissenschaftszentrum Berlin für Sozialforschung during the stay of the first author in Berlin. The SIPP data have kindly been provided by the Bureau of the Census and the ICPSR at the University of Michigan, Ann Arbor. Of course, neither institution is responsible for our analysis of the data, nor our interpretation of the findings.

## **Abstract**

Propensity score matching provides an estimate of the effect of a “treatment” variable on an outcome variable that is largely free of bias arising from an association between treatment status and observable variables. However, matching methods are not robust against “hidden bias” arising from unobserved variables that simultaneously affect assignment to treatment and the outcome variable. One strategy for addressing this problem is the Rosenbaum bounds approach, which allows the analyst to determine how strongly an unmeasured confounding variable must affect selection into treatment in order to undermine the conclusions about causal effects from a matching analysis. Instrumental variables (IV) estimation provides an alternative strategy for the estimation of causal effects, but the method typically reduces the precision of the estimate and has an additional source of uncertainty that derives from the untestable nature of the assumptions of the IV approach. A method of assessing this additional uncertainty is proposed so that the total uncertainty of the IV approach can be compared with the Rosenbaum bounds approach to uncertainty using matching methods. Because the approaches rely on different information and different assumptions, they provide complementary information about causal relationships. The approach is illustrated via an analysis of the impact of unemployment insurance on the timing of reemployment, the post-unemployment wage, and the probability of relocation, using data from several panels of the Survey of Income and Program Participation (SIPP).

# **Assessing Bias in the Estimation of Causal Effects: Rosenbaum**

## **Bounds on Matching Estimators and Instrumental Variables**

### **Estimation with Imperfect Instruments**

#### **Introduction**

In recent years, the “counterfactual” approach to causal analysis has made important inroads into statistical and econometric work on causal inference (e.g. Holland 1986, Rosenbaum 2002; Heckman, LaLonde and Smith 2000), and has entered sociological research through a series of papers that articulate the major differences between this new approach and the standard regression approach to causal inference (Sobel 1995, 1996; Winship and Morgan 1999).

Arguably the empirical strategy from the new causal analysis literature that has attracted the greatest attention is the method of matching, and specifically the propensity score approach to the method of matching. Propensity score matching is a method that arguably improves on the ability of regression to generate accurate causal estimates by virtue of its nonparametric approach to the balancing of covariates between the “treatment” and the “control” group, which removes bias due to observable variables.

However, matching methods are not robust against “hidden bias” arising from the existence of unobserved variables that simultaneously affect assignment to treatment and the outcome variable. One strategy for addressing this problem is the Rosenbaum bounds approach, which allows the analyst to determine how strongly an unmeasured confounding variable must affect selection into treatment in order to undermine the implications of a matching analysis. Instrumental variables estimation provides an alternative strategy for consistent estimation of causal effects, but the method typically reduces the precision of the estimate and has an

additional source of uncertainty that derives from the untestable assumptions of the instrumental variables (IV) approach.

This paper makes two principal contributions to the new literature on causal analysis. First, we discuss and implement two versions of the Rosenbaum bounds strategy for assessing the potential impact of hidden bias. Second, we formalize the uncertainty in IV estimation for an important subset of IV estimators in a manner that is parallel to the approach used by Rosenbaum. Because the IV and Rosenbaum approaches rely on different information and different assumptions, they provide complementary information about the potential causal relationship between the treatment and outcome variables in question. To illustrate the approach, we compare the utility of these alternative strategies for estimating the impact of unemployment insurance on three outcomes: the timing of reemployment, the post-unemployment wage, and the probability of relocation. STATA code for computing these sensitivity analyses is available on the referenced website.<sup>1</sup>

## **Estimation of Causal Effects in the Presence of Heterogeneity**

The counterfactual concept of causality is increasingly familiar to social scientists, and so we develop here only the concepts that are needed in our subsequent development of the topic of bias assessment. The standard regression framework makes the implicit assumption that causal effects are constant in the population. Population heterogeneity of causal effects is typically addressed in the standard approach via the inclusion of interaction effects, which are intended to capture the systematic variation of an effect with the value of some other observable variable. Whether or not such interactions are included, however, the standard regression approach implicitly assumes that causal effects are constant either in the population or within a subpopulation defined by the interaction variables. While regression estimates are sometimes

informally treated as averages of heterogeneous causal effects, the definition of an average causal effect has been made much more precise in the recent statistical literature on causal estimation.

In a series of papers, a set of researchers including Heckman and coauthors (Heckman and Vytlacil 2002), and Angrist and coauthors (Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996; Angrist and Krueger 2001) have distinguished several distinct causal effects, including (a) the average treatment effect (ATE), (b) the average treatment effect for the treated (ATT),<sup>2</sup> and (c) the local average treatment effect (LATE), which is defined with respect to some specific instrumental variable. While the literature does not use the terminology, we think it useful for present purposes to distinguish further between “unconditional” and “conditional” average treatment effects, and unconditional and conditional average treatment effects for the treated. By “unconditional average treatment effect,” we mean the average effect of the treatment in the population. By “conditional average treatment effect,” we mean the average effect of the treatment for some subpopulation that can be defined by observable variables. The unconditional and conditional ATT can be similarly defined. We use the modifiers “conditional” and “unconditional” in the text below to distinguish the two sets of estimators; when ATT or ATE are unmodified, we are referring to the conditional ATT or ATE.<sup>3</sup> How these quantities differ depends upon the specific mechanism that generates the heterogeneity in responses to treatment.

Using the terminology from Heckman (1997), one can write the following nonparametric model for the treatment effect.

$$\begin{aligned} Y_{0i} &= \mu_0(X_i) + U_{0i} \\ Y_{1i} &= \mu_1(X_i) + U_{1i} \end{aligned} \tag{1}$$

where  $Y_0$  and  $Y_1$  are the corresponding outcomes for unit “i” according to whether this unit receives the treatment ( $D_i = 1$ ) or not. Only one of these outcomes is observable for each unit.  $X$

corresponds to all observed covariates that have a structural effect on the outcome. The effect of these variables on the outcome will generally depend on whether or not the unit is treated (i.e.,  $\mu_1(X) \neq \mu_0(X)$  in general).  $U_0$  is the effect of unobservable variables if there is no treatment, and  $U_1$  is the effect of unobservable variables if there is a treatment for each unit.

Using this notation, one can distinguish the following quantities:

$$E(\Delta | X) = \mu_1(X) - \mu_0(X) \quad (2)$$

$$E(\Delta) = \int_X [\mu_1(X) - \mu_0(X)] d(P(X)) \approx \sum_X E(\Delta | X) p(X) \quad (3)$$

$$E(\Delta | X, D = 1) = \mu_1(X) - \mu_0(X) + E(U_1 - U_0 | X, D = 1) \quad (4)$$

$$\begin{aligned} E(\Delta | D = 1) &= \int_X [\mu_1(X) - \mu_0(X) + E(U_1 - U_0 | X, D = 1)] d(P(X)) \\ &\approx \sum_X E(\Delta | X, D = 1) + E(U_1 - U_0 | X, D = 1) p(X) \end{aligned} \quad (5)$$

Here  $E(\Delta | X)$  is the average treatment effect, conditional on X,  $E(\Delta)$  is the unconditional average treatment effect,  $E(\Delta | X, D = 1)$  is the average treatment effect on the treated, conditional on X, and  $E(\Delta | D = 1)$  is the average treatment effect on the treated, unconditional on X. These equations imply that the effect of the treatment will typically vary from person to person, because of inter-personal variation in X. They further imply that the treatment will vary from person to person because of the effects of unobservable variables in  $U_0$  and  $U_1$ .

As Heckman has further shown, one can rewrite the two equations in (1) in either of the following ways:

$$Y = \mu_0(X) + D[E(\Delta | X)] + \{U_0 + D(U_1 - U_0)\} \quad (6)$$

and

$$Y = \mu_0(X) + E(\Delta | X, D = 1) + \{U_0 + D[U_1 - U_0 - E(U_1 - U_0 | X, D = 1)]\} \quad (7)$$

These equations make clear the distinction between the ATE and the ATT. These equations also demonstrate the conditions under which the ATE equals the ATT, and the conditions under which estimators for these two quantities will be consistent. The ATE = ATT when

$$E(U_1 - U_0 | D = 1) = 0$$

i.e., when the impact of unobserved variables on outcomes is expected to be the same regardless of whether the units are treated, conditional on the units themselves being treated. This condition would be violated whenever individuals have private knowledge of these potential outcomes, and are more likely to choose to be treated when the comparison of the “return” to these unmeasured variables is more favorable in the state of being treated. The condition is satisfied if the unmeasured variables have the same impact in either state, or when there is no relationship between differences in these potential effects and the assignment process.

The statistical problem that needs to be solved in any specific scientific study is the construction of an estimator for some well-defined average causal effect. As noted above, standard regression frameworks do not rigorously define any notion of an average treatment effect. An alternative approach, namely the random coefficients statistical framework (which includes multilevel models estimated by such programs as HLM or MLn) overcomes this limitation; the estimated coefficients of a random coefficients model estimates in principle the mean of the population distributions of the causal effect of interest. But several of the assumptions that underlie the random effects estimator are very strong. One strong assumption is that the precise form of the effect of potentially confounding variables on the dependent variable has been captured by the model specification. A second strong assumption is that the model specification correctly extrapolates to areas of the population distribution that are not well

“supported” by the sample of treated units. The invalidity of either of these assumptions biases the estimate of the population average causal effect.

Matching methods, of which propensity score matching is an important subset, are attractive because they do not rely for their validity on either of these two assumptions. Matching methods focus attention on a specific causal effect of interest, and treat all variables other than the treatment variable as potentially confounding variables. In the matching approach, the influence of confounding variables is reduced by the method of covariate balance, i.e., by matching the potentially confounding covariates of the cases that experienced the treatment with cases that did not experience the treatment. A perfect matching (whether on the individual covariates or on the propensity score) eliminates any relationship between the covariates and assignment to the treatment, and hence eliminates the possibility of bias from these variables.

However, matching methods still have an important limitation that they share with random coefficients regression. Both methods require that the assignment to treatment be “ignorable” given the observed covariates (Rubin 1978). In other words, they require that, conditional on the observed covariates, the process by which units are selected into treatment be unrelated to unmeasured variables that affect the outcome variable. In light of the distinction between causal effects made above, the assumption of ignorability can be further explained as follows. Conditional on  $X$  and assuming that the assignment process was independent of  $U_0$ , matching methods estimate what we referred to earlier as the conditional ATT.<sup>4</sup> Conditional on  $X$  and assuming that the assignment process was independent of both  $U_0$  and  $U_1$ , then the  $ATT=A TE$  and matching methods would consistently estimate the conditional ATE, which would equal the conditional ATT. Assuming that the assignment mechanism is independent of both  $U_0$  and  $U_1$ , then the estimated causal effect obtained from the application of matching

methods to the entire matched sample is a consistent estimator for the unconditional ATT. An estimate of the unconditional ATE effect can be obtained via a re-weighting of the heterogeneous treatment effects estimated at different levels of  $X$  or (equivalently) at different values of the propensity score. If the assignment mechanism is only independent of  $U_0$ , then the estimated causal effect obtained from the application of matching methods to the entire matched sample is a consistent estimator for the unconditional ATT, but a reweighting of the data would not in general give the unconditional ATE.

The assumption of ignorability is very strong. In the typical analysis of data that is non-experimentally collected, the assignment mechanism will not be independent of  $U_0$  or  $U_1$ . Under these circumstances, the estimate obtained from matching methods will be biased. The question for empirical research then becomes an assessment of the utility of the matching estimator in the presence of this bias and its performance against alternative estimators that explicitly attempt to correct for this bias. Two complementary strategies are discussed and compared in the next section of this paper, namely the use of Rosenbaum bounds and the use of instrumental variables estimators with sensitivity analysis for the validity of the instruments.

## **Bias Correction and Bias Assessment using Matching Methods and IV Estimation**

Because social science data is typically “observational” in character (i.e., the data are not collected via an experiment), the process by which individuals are assigned or assign themselves to treatment is not typically ignorable, and the treatment variable is typically endogenous. Perhaps the most common strategy for dealing with this problem in the social sciences is the method of instrumental variables (IV) estimation. Less well-known is the Angrist, Imbens and Rubin (1996) interpretation of this estimate as the “local average treatment effect.” Their

reinterpretation follows directly from the recognition that the treatment effect is typically heterogeneous in the population even after observable variables are controlled.

As Angrist and colleagues have shown, the standard interpretation of the IV estimation as an estimator for the treatment effect only applies under the unrealistic case when the treatment effect is constant within the population, which further implies that the conditional ATE equals the unconditional ATE, the conditional ATT and the unconditional ATT. In the more realistic case shown in equation (1), and under a set of additional assumptions,<sup>5</sup> the IV estimator estimates the “local average treatment effect,” or “LATE,” which is the average effect of the treatment for the subsample of the population which is induced by a specific change in the value of the IV to select themselves into treatment.

The LATE estimator, whose asymptotic distribution and variance is described in Imbens and Angrist (1994), is defined with respect to a specific IV. The presumption is that the analyst can model the process of assignment to treatment as a function of an IV (referred to here as  $Z$ ). IV estimation requires that the assignment mechanism, conditional on  $Z$ , is ignorable. By this is meant that  $Z$  affects assignment to the treatment  $D$ , and that, after controlling for  $X$ ,  $Z$  is not related to any unobservable factors that affect the outcome. Angrist, Imbens and Rubin (1996) have shown that the LATE estimator can be conceptualized as the ratio of two matching estimators. The numerator gives an estimate of the impact of a change in the IV on the outcome, while the denominator gives an estimate of the impact of a change in the IV on the probability of treatment.

An IV with only two possible values generates a single LATE estimator. An IV with multiple values can generate multiple LATE estimators, each of which is defined with respect to a specific change in the value of the IV. If the domain of the IV consists of a discrete set of

values, and if the IV has a monotonic relationship to the probability of treatment (one of the conditions for the IV estimator to be a LATE), then the standard IV estimator equals a weighted combination of LATE estimators defined with respect to each successive increase in the value of the IV (Imbens and Angrist 1994). Monotonicity is satisfied when the natural metric of the IV has a monotonic relationship to the probability of selection into treatment. Monotonicity is also satisfied if each discrete point of the IV is scored as equal to the propensity score (i.e., the probability of selection into treatment).

The properties of the IV/LATE estimator define the uncertainty contained in the LATE for the causal effect that it estimates. The most obvious source of uncertainty is contained in the confidence interval of the LATE estimator. Because the LATE is a ratio of two matching estimators (for the effect of Z on Y and for the effect of Z on D), its variance is generally larger than is the variance of the matching estimator for the effect of D on Y. However, the estimated confidence interval is only one component of this uncertainty. The LATE estimator relies for its consistency on the assumption that the assignment mechanism (based on Z) is ignorable (i.e., that Z is uncorrelated with unmeasured factors that affect Y, after conditioning on observable variables). Uncertainty about the validity of this assumption implies additional uncertainty about the true causal effect beyond that contained in the estimated confidence interval. Uncertainty about the validity of other assumptions (such as SUTVA and monotonicity) must also be evaluated in any specific application before the estimate can be interpreted as a LATE or (even more strongly) as an ATE or ATT estimator (Angrist, Imbens, and Rubin 1996; Angrist 2003).

An alternative approach to the assessment of uncertainty in causal estimation has recently been suggested by Rosenbaum (2002). Rosenbaum's approach does not rely on the search for variables that putatively satisfy the assumption for IV estimation of the LATE, ATE, or ATT. Instead, his approach begins with estimating the ATT using matching methods based on the

strong assumption of ignorable assignment, conditional on  $X$ . In the second step, one postulates the existence of a confounding variable  $W$ , which affects the odds of being assigned to the treatment (i.e., the odds that  $D=1$ ), conditional on  $X$ . In the trivial case where  $W$  is uninformative about  $D$ , then the assignment process is random and the Rosenbaum bounds equal the significance level estimated by the matching estimator. As the potential impact of  $W$  on  $D$  (expressed in terms of odds ratios) is assumed to be stronger, the confidence interval on the estimated effect becomes wider, and the significance level of the test of the null hypothesis of no effect of  $D$  on  $Y$  increases (i.e., the p-value goes up). Third, for each assumed level of association between  $W$  and  $D$ , one determines the end points on the bounds for the significance level of the test of the null hypothesis for the case where  $W$ 's effect on the outcome is so strong that knowledge of  $W$  would perfectly predict which of a pair of matched cases would have the higher response regardless of which case received the treatment. Uncertainty about the true effect of the treatment is therefore expressible in two dimensions of uncertainty, i.e., (1) the confidence interval on the estimated treatment effect (or equivalently, the significance level of the test of the null hypothesis of no effect), which is wider (the p-value is large) as the assumed relationship between  $W$  and  $D$  grows stronger, and (2) uncertainty about how big is the relationship of  $W$  and  $D$  under the assumption that  $W$  has a strong effect on  $Y$  (i.e.,  $W$  can determine with high probability the relative outcomes of matched cases). We describe his procedure more fully in appendix A.<sup>6</sup>

In typical applications of the IV approach, the analyst assumes that the proposed  $Z$  satisfies the required assumptions and asserts that the IV estimator is superior to an OLS or a matching approach because it has taken account of the endogeneity arising from  $W$ . Because typical applications assume away one of the major uncertainties inherent to the approach (namely the uncertainty about the validity of  $Z$  as an instrument), one might refer to these typical

approaches as “naïve” IV estimation. The Rosenbaum bounds approach provides an alternative approach to the endogeneity problem by making explicit the extent to which the assumptions underlying “naïve” applications of the matching estimator fall short. If  $Z$  is a valid instrument, the IV estimator based on this  $Z$  provides more precise information about the effect of  $D$  on  $Y$  to the extent that the confidence interval of the IV estimator is tighter than the (widened) confidence interval of the matching estimator given by the Rosenbaum bounds under reasonable guesses about the impact of  $W$  on selection, i.e., reasonable guesses about  $\gamma$  in equation (12) or about  $\Gamma$  in equation (13) of Appendix A. If, however, the bounds of what we have termed the naïve IV estimator are wide relative to the Rosenbaum bounds, given reasonable values of  $\gamma$ , then IV with the specific  $Z$  in question provides less information about the effect of  $D$  on  $Y$  than do matching methods even if  $Z$  is a valid instrument (making the IV estimator consistent) and even if the matching estimator is biased by a nonignorable treatment assignment process.

In practice of course, the analyst does not know with certainty whether  $Z$  is a valid instrument any more than the analyst knows the value of  $\gamma$  with certainty. To increase the information content of IV estimation, we suggest that the uncertainty about  $Z$  be formalized and the impacts of this uncertainty computed in a fashion parallel to that used in the Rosenbaum bounds approach. Because the uncertainty embedded in IV estimation is distinct from the uncertainty embedded in the matching approach, the use of both methods simultaneously can increase the overall information available to the analyst about the effect of  $D$  on  $Y$ .

Establishing bounds on the IV estimator to reflect uncertainty about the validity of  $Z$  as an instrumental variable is most easily done by conceptualizing the bias as a form of omitted variable bias. While the omitted variable approach to bias in IV estimation is quite general, we develop the approach in this paper for the special (but quite frequent) case where  $Z$  is a “grouping

variable.” A grouping variable, which can also be described as a “macro-variable,” is any variable that divides the sample into a set of disjoint groups. Since potential instrumental variables are often grouping variables, this special case is highly relevant to many empirical applications. One well-known example of a grouping variable comes from approaches to estimating the effect of military service on career outcomes during the Vietnam war era that use sample member’s draft lottery numbers as instrumental variables (Angrist, Imbens and Rubin 1996; Moffitt 1996b). Studies of the effects of a social policy on outcomes often use geographic location as a grouping instrumental variable in cases where the policy has not been implemented everywhere (thus creating a “natural experiment”). In the case of the draft lottery number, the value of the IV could be the lottery number itself if this was monotonically related to the probability of military service. For community residence, the IV could be scored as the probability of selection into treatment for the particular community, or it could consist of a set of dummy variables for community of residence. For the grouping variable to be a valid IV, it must be the case that one’s group location affects the probability of treatment, but that group location itself does not affect the outcome, and is not related to unmeasured variables that affect the outcome variable,

As Moffitt (1996a, 1996b) has shown, IV estimation using a grouping variable as the IV variable is identical to a regression of group means for the outcome variable on the treatment variable. To be precise, let us assume that we condition on observed covariates, or equivalently on the propensity score  $P(X)$ . For each value of  $X$  or of  $P(X)$ , we specify that

$$\begin{aligned} y_i &= \mu_0(X_i) + D\delta(X_i) + \varepsilon_i \\ &= \mu_0(X_i) + D\delta(P(X_i)) + \varepsilon_i \end{aligned}$$

where  $\delta(X)$  is the coefficient of the treatment variable, and  $P(X)$  is the propensity score.

Assume that there is some instrumental variable  $Z$  that classifies the individual units into  $J$  groups

(we will give examples later). Then the IV estimator for  $\delta(X)$  is equivalent to regressing the average  $y$  within each of the  $J$  groups of all observations with the same value of  $X$ , or equivalently, the same  $P(X)$ , i.e.,

$$\bar{y}_j(P(X)) = \bar{D}_j(P(X)) * \delta_L(P(X)) + \bar{\varepsilon}_j(P(X)) \quad (8)$$

If  $J=2$ , then  $\delta_L$  is the LATE and the  $\hat{\delta}_L$  estimated by OLS on the aggregate data is an estimate of the LATE. If  $J$  is greater than 2, then  $\hat{\delta}_L$  is the weighted LATE estimator of Imbens and Angrist (1994).

With this framework, we are now in a position to address the question of IV bias. The most common assumed reason for the failure of the IV assumptions concerns the failure of what Angrist, Imbens and Rubin (1996) refer to as the exclusion assumption, namely, the assumption that  $Z$  has no structural effect on  $y$ , and that the value of  $y$  is independent of  $Z$ , given  $D$  and  $X$ . This assumption is violated when  $Z$  is correlated with  $\varepsilon$ . When  $Z$  is a grouping variable, the failure of the exclusion assumption implies that (given  $X$ ) the average value of  $\varepsilon$  varies across the  $J$  groups. We write this failure in terms of a group level characteristic  $W_j$ , so that we can write

$$\varepsilon_j = \gamma W_j + v_j, \quad (9)$$

where  $v$  is assumed to be independent of  $D$ . Under this formulation, the bias of the IV estimator can be expressed in terms of the standard formula for omitted variable bias, namely

$$E(\hat{\delta}_L | P(X)) = \delta_L + \alpha\gamma \quad (10)$$

where  $\alpha$  is the slope parameter from a regression of  $W$  on  $\bar{D}$ .

Given this formulation, it now becomes possible to express the additional uncertainty of the LATE estimator in terms of assumptions about the size of the effect of  $W$  on  $\bar{y}$  and about the effect of  $W$  on  $\bar{D}$ . By combining plausible guesses about the size of the effect of  $W$  on  $\bar{y}$  in the aggregate equation (8) and in the regression of  $W$  on  $\bar{D}$ , one can conduct the same type of sensitivity analysis on the IV estimator that one can do on the matching estimator with the Rosenbaum bounds approach.

For both theoretical and practical reasons, we typically do not assume that the treatment effect varies with all observed covariates. A “main effects” version of the above model would include  $X$  as a set of covariates in the model. Then the aggregation version of IV is obtained in three steps. In the first step, one obtains the residual of  $y$  from a regression of  $y$  on the  $X$  variables. In the second step, one obtains the residual of  $D$  from a regression of  $D$  on the  $X$  variables. In the third step, one computes the mean  $y$ -residual within each of the  $J$  groups, and regresses that on the mean  $D$ -residual within each of the  $J$  groups. The result is the IV estimator of  $y$  on  $D$  when grouping variable  $Z$  is the instrumental variable. Given this formulation, bias in the IV estimator can be expressed in terms of the omitted group-level characteristic  $W$  as above.

In summary, both IV and the Rosenbaum bounds methods approach can be used to address possible bias in OLS or matching estimators. Both approaches necessarily widen the confidence interval around the estimate of the causal effect. With the Rosenbaum bounds sensitivity analysis for the matching estimator, the two additional sources of uncertainty are first the amount of bias in the treatment assignment, and second, the impact of this bias on the bounds for the treatment effect. With the IV estimator, the two additional sources of uncertainty are first the wider confidence interval on the IV estimator relative to the OLS or matching estimator, and second the potential bias in the IV estimate due to a failure of the assumptions underlying this

estimator.<sup>7</sup> We have described the procedure proposed by Rosenbaum to measure uncertainty with the matching estimator and have proposed a parallel procedure based on the assumption that the IV estimator in question is a grouping variable and that the major uncertainty concerns the validity of the exclusion assumption. The difference between the two procedures is that the Rosenbaum approach provides an assessment of the uncertainty in the ATT, where the ATT is either unconditional or conditional depending upon whether the matching is done across the entire sample of treated individuals or on a subsample defined with respect to  $X$ . The IV sensitivity analysis provides an assessment of the uncertainty of the IV estimator, which equals the LATE or the weighted LATE<sup>8</sup> under a set of additional assumptions such as monotonicity, and which equals the ATT or the ATE under even stronger assumptions.

In the next section, we compare the information provided by these two approaches an empirical illustration. Our illustration concerns the behavioral effects of unemployment insurance, i.e. the consequences of welfare state institutions providing financial resources to compensate workers' earnings losses during unemployment spells. We examine three kinds of behavioral responses to unemployment benefits: the first concerns the effect of receiving unemployment benefits on the duration of unemployment. The second concerns the effect of receiving unemployment benefits on the post-unemployment wage. The third concerns the effect of receiving unemployment benefits on the probability of moving to a different state within roughly 18 months following the onset of unemployment. These examples are chosen to illustrate the methods both for continuous and discrete outcomes. The standard predictions (cf. Mortensen 1986)) are that unemployment benefits prolongs unemployment duration, increase post-unemployment wages and reduce the likelihood that workers are regionally mobile, and so these predictions constitute the working hypotheses for our illustration.

## **An Empirical Analysis: The Effects of Unemployment Benefits**

We test the three hypotheses using data on employment histories and geographical mobility for a sample of previously employed workers in the combined 1984, 1986, 1988, 1990, 1992 and 1993 panels of the Survey of Income and Program Participation (SIPP).<sup>9</sup> With a 12-year observation window from January 1984 to December 1995, these combined SIPP panels yield a sample of 24,100 unemployment spells experienced by 21,551 workers. We measure unemployment duration as the probability that an individual exited unemployment within the first three months of an unemployment spell, where the exit could occur either through entering a job or leaving the labor force entirely. We use the log of the change in gross real wages between a workers' last job immediately prior to the unemployment spell and first job after leaving unemployment as the second outcome variable.<sup>10</sup> We measure regional mobility as the probability that an individual lived in a different state 18 months after the onset of the unemployment spell.

The treatment variable is the receipt of unemployment benefits. To estimate the effect of the receipt of benefits on the three outcomes, we first use propensity score matching, where we contrast individuals who received benefits at any time during the unemployment spell with observationally similar individuals who did not receive benefits. We estimate the effects of benefit status on outcomes. Our second strategy allows for the possibility that individual benefit status is endogenous to the outcomes of interest, and looks for plausible exogenous instruments for individual benefit status in order to estimate the causal effects of unemployment benefit status on the outcomes via IV estimators. Endogeneity would occur if some of the factors that influence receipt of benefits also affect the duration of unemployment, the wage change following return to work, or geographic mobility.

One plausible instrumental variable strategy is to control for characteristics of the labor market, which likely would affect unemployment duration, wage change, and geographical mobility, and to use state-level variation in UI systems as the instrument. State-level UI coverage would qualify as an instrument if (1) an individual's state location was random, or at least unrelated to the outcome variables, conditional on  $X$ , (2) state-level UI coverage affected the probability of an individual's receiving unemployment insurance, and (3) the effect of state UI coverage on outcomes occurs only through the effect of an individual's UI coverage on outcomes. If the monotonicity assumption was also true (i.e., no individual would be more likely to receive unemployment insurance by virtue of a lowering of UI coverage in the state of residence), then the use of state-level UI as an instrument would produce a set of LATE estimates across the spectrum of observed variation in state UI coverage.<sup>11</sup>

The ability of state-level UI coverage to meet the above conditions clearly depends upon the comprehensiveness of  $X$ . In general, one would expect that state location is not random and not ignorable with respect to the outcomes. But state location becomes ignorable if all factors aside from UI that determine state location which are related to the outcome variables are in  $X$ . Assumption (2) is clearly satisfied. Condition (3) in this case amounts to the SUTVA, i.e., it implies that treatment or lack of treatment of others in the state has no impact on an individual's outcome apart from that individual's own treatment status. The monotonicity assumption implies that changes in UI coverage occur only through mechanisms that equally affect all unemployed in the state. If, for example, a state raised UI coverage by tightening one eligibility rule while loosening another, then the monotonicity assumption would be violated. Because the validity of state UI coverage as an instrument depends upon the comprehensiveness of the information in  $X$ , we consider state UI coverage to be a typical (and typically imperfect) instrumental variable for unemployment insurance.

### *Sensitivity Analysis via the Rosenbaum Bounds Method of Matching Estimators*

To obtain the propensity score matching estimator, we first use individual benefit status to form matched pairs of observationally similar workers who had received benefits during the unemployment spell (the treatment cases) and workers who did not (the controls). Matching was done on the individual propensity score of receiving benefits. The propensity score was operationalized as the predicted probability of receiving benefits estimated from a logistic regression of unemployment benefit status on pre-unemployment wage, education, labor force experience, tenure with previous employer, gender, race and other predictors. The coefficients from this model, which are presented in Table 1, show that the likelihood of receiving benefits rises with a worker's previous wage, labor force experience, and tenure with the previous employer.

#### TABLE 1 ABOUT HERE

Because unemployed workers who receive unemployment benefits typically had higher pre-unemployment wages and more experience than the unemployed who do not receive benefits, a simple comparison of mean outcomes for sample members who do and who do not benefits is unlikely to yield accurate estimates of the causal effect of the receipt of unemployment benefits on outcomes. By forming matched pairs of observationally similar treatment and control cases, matching methods eliminate the confounding effects of observable variables. In our example, we use a stratified 1x1 random-order, nearest-neighbor caliper matching algorithm that matches treatment and control cases with similar propensity scores within the tolerance level (the caliper) for acceptable matches defined in terms of the empirical variance of the propensity score. Since we have rather large samples of unemployed workers from the SIPP, we can use a fairly strict caliper of  $w = 0.05$ , i.e. we can require a high degree of observational similarity between treatment and control cases in our analysis and still find matching control cases for our treatment

cases. To further increase the comparability of treatment and control cases, we conducted a stratified matching by states and time period. Full details of the matching algorithm are given in Appendix B.

Table 2 demonstrates how matching restricts the control sample in order to increase the similarity of the subsample of control cases that are directly compared with the treated cases in order to estimate the consequences of treatment.<sup>12</sup> Table 2 presents the means for propensity scores and, in the case of the wage model, all other covariates before and after matching. In almost all cases, it is evident that sample differences in the raw data significantly exceed those in the samples of matched cases. The process of matching thus creates a high degree of “covariate balance” between the treatment and control samples that are used in the estimation procedure.

TABLE 2 ABOUT HERE

We use the standardized mean difference between treatment and control samples, i.e.

$$bias = \left| \frac{100 (\bar{x}_T - \bar{x}_C)}{\sqrt{\frac{(s_T^2 + s_C^2)}{2}}} \right|$$

as a convenient way to quantify the bias between treatment and control samples (cf. Rosenbaum and Rubin 1985; Rubin 1991), where  $\bar{x}_T$  and  $s_T^2$  are the sample mean and variance for the treatment subsample, and  $\bar{x}_C$  and  $s_C^2$  are the comparable statistics for the control subsample, and essentially provides a measure of the difference in means for each x in standard deviation units. By this measure, imbalance between treatment and control samples in terms of the propensity score amounts to more than 80% in the raw data for each of the three models (i.e., the difference in propensity scores for the unmatched treatment and control sample is over 80% as large as the

standard deviation). This bias is reduced to levels well below 10% by the matching process, which amounts to more than a 90% reduction in bias for each of our outcome variables.

An examination of covariate balance for the specific covariates in our analysis shows that propensity-score matching tends to put heavy emphasis on achieving covariate balance on the key predictors from the logit model, i.e. in terms of earnings, experience, tenure, gender and in the proportion of black workers in the two samples. In general, there will be a trade-off between the size of the bias reduction and the proportion of the treated cases that can be matched, with the size of the tradeoff depending upon the size of the treated and control samples. (see Appendix Table C for an illustration of this relationship). With the relatively large sample sizes of the SIPP data, we were able to match about 60% of the treatment cases using the stringent caliper of 0.05.

Because propensity-score matching removes most of the bias attributable to observable covariates, we can use the difference in mean outcomes in the matched samples to obtain an estimate of the average treatment effect on the treated. Table 3 gives the estimates of the unconditional ATT of unemployment benefits based on the stratified propensity score matching, and compares these to standard OLS estimates.<sup>13</sup> The first column of Table 3 gives mean outcomes among treatment cases (i.e. workers with benefits), while the second column gives the mean outcomes among all control cases in the sample. The difference between these two quantities is the “naïve” estimate of the unconditional average treatment effect, uncorrected for the possibly confounding effects of observed covariates. The third column shows the mean outcome among the set of matched controls. The average treatment effect of unemployment benefits on the treated workers actually receiving benefits (the unconditional ATT) is given in column 4, and is simply the difference between columns 1 and 3. Finally, column 5 shows the OLS estimate for the effect of unemployment benefits on each of the three outcome variables.

### TABLE 3 ABOUT HERE

Table 3 shows that the estimated effect from propensity score matching supports all three hypotheses on the effects of unemployment benefits: Workers who had access to benefits achieve higher post-unemployment wages, exit unemployment less quickly and are less likely to move to another state. Compared to the naïve estimator, propensity-score matching resulted in larger estimated treatment effect estimates for all three outcomes. In the case of wage change, the naïve estimator would have given a result that not only had a different magnitude but even had a different sign than the propensity-score matching estimate. Compared to the OLS estimates of column 5, the matching estimates tend to be slightly more conservative (i.e. lower) and estimated less precisely (i.e. with larger standard errors). There can actually be two views on this latter aspect: on the one hand, the larger standard errors of matching estimates can be seen as a consequence of matching being a data-intensive technique that discards information contained in the non-matched cases. On the other hand, since matching is a non-parametric estimator based on samples that exhibit common support, the higher precision of OLS (or parametric methods more generally) can be seen as resulting from untested assumptions in terms of functional form, or, equivalently, the higher standard errors of the matching estimates relative to OLS convey the level of uncertainty of the estimate that can be achieved when one is unwilling to make the parametric assumptions of OLS.

Propensity-score matching estimators are not consistent estimators for treatment effects if the assignment to treatment is endogenous, i.e., if unobserved variables that affect the assignment process are also related to the outcomes. In order to estimate the extent to which such "selection on unobservables" may bias our qualitative and quantitative inferences about the effects of unemployment benefits, we present the results from using Rosenbaum's (2002) procedure for bounding the treatment effect estimates in Table 4. There we give the results of the p-value from

Wilcoxon sign-rank tests for the averaged treatment effect on the treated while setting the level of hidden bias to a certain value  $\Gamma$ , which--as described in more detail in appendix A--reflects our assumption about unmeasured heterogeneity or endogeneity in treatment assignment expressed in terms of the odds ratio of differential treatment assignment due to an unobserved covariate. At each  $\Gamma$  we calculate a hypothetical significance level “p-critical”, which represents the bound on the significance level of the treatment effect in the case of endogenous self-selection into treatment status.<sup>14</sup> By comparing the Rosenbaum bounds on treatment effects at different levels of  $\Gamma$  we can assess the strength such unmeasured influences would require in order that the estimated treatment effects from propensity score matching would have arisen purely through selection effects.

#### TABLE 4 ABOUT HERE

Table 4 shows that robustness to hidden bias varies considerably across the three outcome variables. The finding of a positive effect of UI on post-unemployment wages is the least robust to the possible presence of selection bias. The critical level of  $\Gamma$  at which we would have to question our conclusion of a positive effect is between 1.10 and 1.15, i.e. is attained if an unobserved covariate caused the odds ratio of treatment assignment to differ between treatment and control cases by a factor of about 1.15. For the regional mobility model it would require a hidden bias of  $\Gamma$  between 1.5 and 1.6 to render spurious the conclusion of a negative benefit effect on mobility. In the case of UI effects on unemployment duration, endogenous self-selection would have to attain values for  $\Gamma$  of between 2.2 and 2.3.

It is important to recognize that these results are worst case scenarios. A value for  $\Gamma$  of 1.15 does not mean that there is no true positive effect of UI on post-unemployment wages. This result means that the confidence interval for the post-unemployment wage effect would include

zero if an unobserved variable caused the odds ratio of treatment assignment to differ between treatment and control groups by 1.15 *and if this variable's effect on post-unemployment wages was so strong as to almost perfectly determine whether the post-unemployment wage would be bigger for the treatment or the control case in each pair of matched cases in the data.* In the case where a confounding variable had an equally strong effect on assignment but only a weak effect on the outcome variable, the confidence interval for post-unemployment wages would not contain zero. To repeat, the Rosenbaum bounds are in this sense a “worst-case” scenario. Nonetheless, they convey important information about the level of uncertainty contained in matching estimators by showing just how large the influence of a confounding variable must be to undermine the conclusions of a matching analysis. So for example, an unobserved variable that perfectly predicted the rank ordering of treatment or control cases in matched pairs and that had a  $\Gamma$  of 1.15 would still not be powerful enough to produce the observed mean difference in unemployment duration for the treatment and the control case. If the confounding effects are only this large, the results of the matching analysis imply that receipt of benefits actually do increase unemployment duration.

To illustrate the magnitude of hidden bias that would cause us to revise our findings of causal effects of unemployment benefits on these three variables, we equate the magnitude of hidden bias expressed by specific levels of  $\Gamma$  in terms of the equivalent effects of observed covariates for which we know the impact on assignment to treatment from our propensity score model. The critical level of  $\Gamma=1.15$  is attained at a difference in log previous wages of more than 0.2 (or more than two dollars per hour for the average worker), or at a difference of 2.7 years of experience, or at a difference of about two years of tenure with a worker's previous employer. Hence, according to the bound estimates, we would have reason to doubt our finding of a causal effect of UI on post-unemployment earnings if we had reason to believe there was an outside

(unobserved) covariate affecting treatment assignment of at least this magnitude that was equivalent to the effect on assignment of \$2 per hour, or 2.7 years of labor force experience, or 2 years of tenure with the previous employer.

The strength of hidden bias required to alter the qualitative conclusions about the effects of unemployment benefits on unemployment duration or regional mobility is in both instances greater than for the post-unemployment wages outcome. For example, the critical  $\Gamma$  of about 1.5 in the case of regional mobility is equivalent to the measured net effect of being black instead of white ( $\exp(b) = 1.48$ ), or the effect of an additional 10 years of labor force experience, or the effect of over 90 months of tenure with the previous employer on assignment to treatment status. Even more powerful unobserved variables must be at work to challenge our estimates of the effect of UI on unemployment duration, and – it is important to remember -- such hidden variables must be able to almost perfectly predict the relative outcomes of the matched treatment and control variables. The hidden bias required to change our conclusions about negative benefit effects on unemployment duration is certainly larger than any single effect that we estimated in our propensity score model – effectively, we would need two *independent* sources of bias in the order of 10 years of experience and 7 years of tenure working *together* to produce a level of  $\Gamma = 1.5 * 1.5 = 2.25$  that was of the order required to challenge our conclusion about the benefit effect on unemployment duration.

### *Sensitivity Analysis of Bias in IV Estimation*

Next, we assess the potential information about causal effects in our empirical examples that can be obtained via instrumental variables estimation. In practice, finding credible instruments can be a daunting task, and researchers typically resort to past covariate values (in panel data settings) or use ‘natural experiments’ in terms of institutional differences across

political units and over historical time (e.g. before and after some policy reform). We follow the latter practice here and use institutional variation in unemployment insurance (UI) systems across U.S. states as our instrument of individual benefit status. State variation in UI systems is a promising instrumental variable insofar as institutional differences are plausibly exogenous to individual labor market behavior (i.e. the average unemployed worker will not be able to adapt UI policies to his or her preferences). Also, there is wide variation in terms of benefit disqualification policies, search requirements, benefit levels and other aspects of state UI systems (cf. Vroman 1990, 2001; U.S. Department of Labor 2002) that might have substantial effects on whether or not a given worker with a given work record receives benefits during an unemployment spell.

State of residence would qualify as an instrumental variable if state of residence had no direct impact on these outcomes, if state of residence had no correlation with unmeasured variables that affect these outcomes (the exclusion condition), and if state of residence was correlated with the treatment variable. The correlation between state of residence and the treatment variable is guaranteed by the fact that the proportion of unemployed receiving benefits varies by state. Because these outcome variables are generally affected by individual-level factors such as age, education, prior work experience and family status, state of residence will be correlated with the error unless all individual-level variables that affect the outcomes and that vary by state are explicitly included in the model. Because the outcome variables will also be influenced by labor market conditions, state of residence will still be correlated with the error even after controlling for individual-level variables unless all pertinent characteristics of the labor market are also included explicitly as covariates in the analysis.

Because information about the most important individual-level and labor market determinants of our three outcome variables are collected in surveys, state-level variation is a

plausible instrumental variable. At the same time it is an imperfect instrument, because it is unlikely that all pertinent individual or labor market variables will be included in the survey or the model. Therefore, we use the procedure described above to perform a sensitivity analysis of the estimated treatment effects from IV estimation. If we conceptualize the IV estimation as an OLS regression of the residualized state-averaged outcome variable on the residualized state-averaged treatment variable, then it becomes clear that the bias from IV regression can be conceptualized as an omitted state-level variable that captures the residual state-level effects of omitted individual and labor market variables on the state-level mean outcome variables.

Before proceeding to IV estimation, we first discuss evidence for the plausibility of the claim that state-level variation can function as an instrumental variable. As a simple check for an association between state policies and workers' treatment status, Figure 1 presents a scatterplot of worker coverage rates across the 45 states (or state clusters) distinguished in the SIPP data, additionally broken down by a more specific indicator of states' good cause policies within the UI system (measured simply by the number of good causes for worker quits acknowledged in state UI regulations).<sup>15</sup> The data are raw proportions of workers having received benefits at any point during an unemployment spell averaged across the 1984-1995 period, and hence not adjusted for any differences in the structure of work forces, labor markets or indeed the structure of unemployment across U.S. states. But even at that basic level, the data speak to significant variation in access to unemployment benefits across states, with coverage rates ranging from a low of some 25% in New Mexico, Georgia, Texas or Colorado up to some 50% in some of the New England states. While there might be numerous reasons for the observed variation across states, the overlaid line connecting the mean coverage rates within levels of state good cause policies certainly indicates some association between state policies and observed benefit levels. The relationship is neither fully linear nor perfect, yet it is evident that liberal state policies

concerning benefit disqualification are positively related to the proportion of workers receiving benefits.<sup>16</sup>

#### FIGURE 1 ABOUT HERE

From this starting point, we can generate IV estimates of the effects of unemployment benefits on our three outcome measures. Table 5 presents three different treatment effect estimates, exploiting and derived from three different specifications of purportedly exogenous state-level variation in individual benefit status. The first set of analyses (presented in the left column) \uses quintiles of aggregate coverage rates as the instrument. The second set of analyses use the full set of 45 state dummies distinguished in the SIPP data. The second measure obviously captures all aspects of the differences in institutional policies across states. The problem with this second measure, however, is that some subset of the 45 state dummy variables is necessarily correlated with any residual state-level differences in individual or labor market variables, and thus would not be a valid instrument. For that reason, the third set of analyses uses the above-mentioned indicator of states' good cause policies as the most direct measure of institutional variation relevant to our question, though even this variable may fail to satisfy the exclusion restriction to the extent that the number of good cause policies is correlated with residual state-level individual or labor market variation.

Our instrumental variables estimation was performed with the same set of individual level controls that were used in the estimation of the propensity score above (cf. Table 1) plus time-varying local unemployment rates and year dummies as crude controls for labor market conditions. Arguably, these covariates capture some part of the potential endogeneity problems and hence promise to yield plausible estimates of the causal effect of unemployment benefits on our three outcome dimensions. The use of 3 alternative sets of instruments produces three distinct

estimates of the effect of unemployment benefits for each of our three outcome variables. We present these alternative IV estimates in table 5.

#### TABLE 5 ABOUT HERE

All three sets of instruments support the same substantive conclusions as the propensity score matching and the standard OLS analyses described before. In all IV analyses, we obtain clear evidence that unemployment benefits improve post-unemployment wages, prolong unemployment duration and lower the propensity of workers to relocate. Compared to both the matching and OLS results presented in Table 1 above, two features of the IV estimates seem noteworthy: first, the estimated magnitude of benefit effects considerably exceed those obtained from both matching and OLS. The IV estimates coincide best with matching and OLS estimates in case of unemployment duration, where the effects estimated by IV are in the order of doubling the respective matching and OLS ones. For both post-unemployment wages and the probability of relocation, IV results indicate much stronger benefit effects than either matching and OLS. Also, – as usually is the case since the inferential base is significantly more restricted – the standard errors of the IV estimates are also well above those from both OLS and propensity score matching.

To perform the sensitivity analysis as described above for the IV case, we postulated the existence of an omitted state-level variable (which in principle includes state variation in omitted individual variables and state variation in omitted labor market variables). The formula for bias in equation (10) can be restated as the triple product of the effect of  $W$  on the outcome ( $\gamma$ ) variable, the standard deviation of  $W$ , and the correlation between  $W$  and the group-mean residualized  $D$  ( $r$ ) divided by the standard deviation of group-mean residualized  $D$ . We estimate the impact of this bias by using a “benchmark” omitted variable that equals the difference (rather than the sum, because the predicted effects of its two components are opposite in sign) of the group mean of

education (an influential observable individual variable) plus the state-level unemployment rate (an influential observable labor market variable). In effect, we assume that actual omitted variables are unlikely to have a confounding effect that would greatly exceed the confounding effect from failing to control for educational differences across states and from failing to control for differences in state-level unemployment rates.<sup>17</sup>

There are two unknowns in the formula for omitted variable bias: the standard deviation of  $W$  and its effect on the outcome variable. To approximate the scale of  $W$ , we assume that it can be represented as the standard deviation of this combined variable and use the result as the scale of  $W$ . To provide reasonable guesses about the likely effect of such an omitted covariate  $W$ , we estimated the effect of the two components of the benchmark omitted variable on each of our three outcome variables. The results, which are shown in appendix D, suggest a maximum effect of  $|.003|$  for the log wage-change outcome,  $|.07|$  for the unemployment duration outcome, and  $|.075|$  for the regional mobility outcome. To be conservative, we assume that the effects of the omitted variable equal the largest bivariate effects of the different components. To approximate the correlation between omitted variables and the treatment variable, we computed the correlation in the aggregate analysis between each component of the benchmark omitted variable and the treatment variable, which is  $.035$  in the case of years of education, and  $.227$  in the case of the unemployment rate.

We used the above results to generate a range for plausible values of  $\gamma$  (i.e., between  $|.01|$  and  $|.10|$ ) and we display the sensitivity analysis for values of the correlation between  $W$  and  $D$  that extend from the high end of what we observe with our benchmark variables to very high correlations. For each outcome, we report results using the sign of  $\gamma$  that would (in combination with a positive correlation) produce an upward bias in the estimated effect of the treatment variable, which of course is the direction of bias that we are concerned about.<sup>18</sup>

Table 6 reports the results of this sensitivity analysis that is in the same spirit as the sensitivity analysis using Rosenbaum bounds that was presented in table 4. For each combination of a specific outcome variable and a specific instrumental variable, we present a sensitivity analysis matrix, which shows the variation in the significance levels of the estimated effect of the treatment variable in the presence of bias generated by a combination of various values for  $\gamma$  and for the correlation between W and the group-mean residualized D (i.e.,  $\text{Corr}(W,D)$ ). To conserve space, we present results for two of our three candidate instrumental variables, namely high vs. low generosity states, and state policy concerning good cause exceptions

#### TABLE 6 ABOUT HERE

For the outcome of log wage change, the left panel (for the state benefit rate quintiles quintiles IV) shows that the effect of unemployment benefits on post-unemployment wage change remains statistically significant at the .05 level so long as the effect of the omitted variable on the outcome is below +0.10 and the correlation between the omitted variable and the treatment variable is less than .25. This is clearly the case for our two benchmark omitted variables (both of whom have a  $\gamma$  of less than +.01), and so we can conclude that our IV estimates are robust to endogeneity of the order that would be created by leaving variables as important education and the state unemployment rate out of the equation. The same conclusion in fact applies regardless of whether state benefit-rate quintiles or liberal good cause regulation is the instrumental variable. Note, however (see the right panel of table 6), that the liberal good cause regulation benchmark IV is more sensitive to the possibility of endogeneity bias, because the standard error of the IV estimate using this IV is about twice as large as when high/low generosity states is used to do the IV estimation (see table 5).

For the unemployment duration outcome,  $\gamma$  was -.021 for the state unemployment rate, and the  $\gamma$  for education (.07) was positive, which means that its exclusion from the model would

cause the IV estimate to understate the true effect of benefits on unemployment duration. As the left panel makes clear, the IV-estimated effect of unemployment benefits on unemployment duration remains statistically significant at the  $p < .001$  level even for omitted variables whose values of  $\gamma$  are five times as large as that for the state unemployment rate and whose correlations with the treatment variable are similar. Based on the IV regression with our state benefit-rate quintiles IV, we would therefore conclude that unemployment benefits increase unemployment duration. Note, however, that we could not draw this conclusion with confidence when liberal good cause regulation was the candidate IV, because it is not robust to endogeneity bias on the scale produced by our benchmark omitted variables.

Finally, we turn to the regional mobility outcome variable. Note that if the omitted variable had a positive effect (as education does on mobility – see appendix D) and was positively associated with the receipt of unemployment benefits, then the IV estimate would understate the true effects of benefits on regional mobility. As for the unemployment duration analysis, the potential problem arises from omitted variables that have the same pattern of effects as the state unemployment rate. The left panel of table 7 shows that a confounding omitted variable would have to have an effect on regional mobility that was three times as large as that of the state unemployment rate while having a correlation with unemployment benefits of commensurate size as the state unemployment rate in order to undermine the IV estimation result. Thus, we again see a basis for reasonable confidence that the IV estimated effect of benefits on regional mobility describes a true causal effect.

## **Comparing the Rosenbaum Bounds and IV Sensitivity Analyses**

The Rosenbaum bounds approach and IV estimation are complementary approaches, not competing ones. It is useful, therefore, to evaluate the different information they give about

causal estimation in our specific example. In this case, both the Rosenbaum Bounds and the IV sensitivity analyses provided important information about the possible causal relationships between our example treatment and outcome variables. Even though propensity score matching by itself does not address the problem of endogeneity due to omitted confounding variables, the Rosenbaum bounds sensitivity analysis provided strong grounds for concluding that unemployment benefits do extend unemployment duration. The sensitivity analysis also provided a reasonably firm basis for concluding that unemployment benefits reduce inter-state mobility. Finally, the sensitivity analysis suggested strong caution in concluding that unemployment benefits increase post-unemployment wages, because the hypothesis test supporting this conclusion is undone even by moderately strong confounding variables so long as such confounding variables have a very strong impact on post-unemployment wages.

IV regression with our three example instrumental variables supported the results of propensity score matching in yielding positive effects of unemployment benefits on all three outcomes. The sensitivity analyses show our IV estimates to be reasonably robust against the presence of bias due to  $W$ , although there are more indications of potential sensitivity in the case of the (more specific) instrument of states' good cause policies than for the instrument using state benefit-rate quintiles, which represent the net impact of state unemployment policies and practices rather than the impact of a single regulation. Our inferences about the effect of unemployment benefits on unemployment duration are the most robust to endogeneity bias. For both post-unemployment wages and regional mobility, however, the sensitivity analyses indicate we might want to be cautious in our conclusions about benefit effects if we have reasons to believe in the presence of an omitted covariate  $W$  like years of education that was very strongly correlated with  $D$  and had about the impact (in absolute magnitude) of years of education on either outcome. In general, and as to be expected, our conclusions about the causal effect of

interest become less secure the higher the correlation between  $W$  and  $D$ , and the stronger the partial effect of  $W$  on the outcome of interest. However, our inferences about causal effects are robust to omitted variables that have roughly both the same magnitude of effect on the outcome variable and correlation with the treatment variable that we observe in the two candidate omitted variables. The fact that the Rosenbaum bounds analysis also provides support for the existence of at least a causal effect on unemployment duration and on inter-state mobility only strengthens our conclusions in this respect.

One important advantage of the Rosenbaum bounds approach is that the underlying estimator (the unconditional or conditional ATT) has a straightforward theoretical interpretation. As we discussed in the theoretical section of this paper, the IV estimates presented in our empirical analysis do not have a straightforward structural interpretation if the effects vary across individuals. Under the assumptions noted above, the IV estimator is actually a weighted combination of LATEs, which does not have clear behavioral implications. In contrast, IV analysis using a policy-relevant binary IV (e.g., whether a state has three as opposed to two good cause exceptions in its unemployment insurance regulations) would provide useful causal information regarding the impact of policy change on change in post-unemployment wages, unemployment durations, and geographic mobility that would work through the impact of policy change on change in benefit coverage. A full analysis would therefore involve a separate computation of the IV estimate for each combination of observed covariate values and at each level of the instrumental variable.

To be concrete, let us take the number of good causes as the proposed instrumental variable and suppose that the effect of unemployment benefits varies by age, which is an observable variable. The LATE could then be computed at different values of age and at different points along the observed variation in the number of good cause exemptions in state

unemployment regulations. The result is a matrix of IV estimates, which provide estimates of the treatment effect for each combination of observed covariates and for a specific policy intervention which increases the number of good cause exceptions by a single unit. The point estimates and standard errors provide information about the extent of variation of the treatment effect along these two dimensions. Because of uncertainty about endogeneity bias for the IV estimates, the above sensitivity analysis could be performed for each element in this matrix of estimates. Because each sensitivity analysis produces a matrix similar to that in table 6, the overall result is a matrix of matrices. The sensitivity results determine whether the researcher can have reasonable confidence in the existence of a measurable change in the state's mean value on the outcome variable from modifying the state's number of good cause exemptions. They also determine whether the researcher can have reasonable confidence that the impact of changing state policies varies according to the current policy or according to age.

If the variation in the estimation matrix is large relative to the variation in the sensitivity matrices, then the result provides useful information about heterogeneity in treatment estimates across individuals and across policy interventions. If the variation in the sensitivity matrices is large relative to the variation in the estimation matrix, then the estimation provides little information about the structure of variation in these estimates. It is possible to conclude that true causal effects exist and that we know something about how they vary across individuals or across policy interventions. Alternatively, we may conclude that true causal effects exist but that we lack sufficient information to understand their heterogeneity. Finally, the combination of point estimates and their standard errors along with the sensitivity analysis may lead one to conclude that there is not sufficient evidence to conclude that a causal effect exists, and *a posteriori*, one can not draw any conclusions about heterogeneity in the causal effect.

To keep this paper within reasonable scope, we do not estimate a full matrix of LATE estimations or perform a sensitivity analysis for each estimate within the estimation matrix that would result from such a full analysis. Given the pattern of results in our Rosenbaum bounds analyses, one might well conclude that it is not necessary for our specific illustrative example to go through such a full set of analyses to draw the conclusion that unconditional average treatment effects on the treated do exist for all three outcome variables. If, however, one was interested in social policy, one might well want to know the impact of raising or lowering state coverage levels on the mean outcomes for all three outcome variables considered in this paper. In such a case, the full set of LATE analyses in conjunction with Rosenbaum bounds analysis of ATT computed at different values of the propensity score would definitely be warranted.

## **Conclusion**

In non-experimental settings, the drawing of firm conclusions about the causal effect of one variable on another inherently involves uncertainty because of the possible confounding influence of other variables. Matching methods provide an effective strategy for controlling the confounding influence of observed variables. However, endogeneity bias is still possible from unobserved variables. Instrumental variable estimation can eliminate endogeneity bias under a set of assumptions that themselves are rather strong and impractical or impossible to verify in most real research settings. This paper has focused instead on strategies for assessing the size of the bias that might affect either matching or IV estimation.

The Rosenbaum bounds approach can often provide reasonable confidence that a causal relationship between a treatment and an outcome variable exists even in the presence of potential confounding variables. The results of this paper support the advice that such analyses be routinely carried out in order to evaluate the robustness of estimates to the possibility of hidden

bias. They also support the use of IV methods while at the same time demonstrating that IV estimation by itself is not a panacea for endogeneity bias. IV estimation does not necessarily provide more information about the “true” effect of a treatment on an outcome variable than is already present in methods such as OLS or matching, which do not explicitly adjust for endogeneity bias. Whether IV provides more or less information about causal effects depends upon how well the candidate instruments satisfy the assumptions underlying IV estimation and upon the width of the confidence intervals of the IV estimates. Because the analyst never knows whether the assumptions underlying IV estimation are valid, we suggest that more attention be paid to the sensitivity of IV estimation to omitted variables that would cause IV estimates to be biased.

Our illustrative results demonstrated the usefulness of both Rosenbaum bounds sensitivity analysis and IV estimation sensitivity analysis in causal estimation. They also demonstrated that these two approaches cannot be given an *in principle* ranking in terms of their information content. In any real situation, the relative information content of the two approaches depends upon the specific relationship being studied and the specific data available to the analyst. The real message of our results is that the two approaches are complimentary. Taken together, they provide more information about the causal processes in question than either method provides by itself.

The information provided by both methods is, it must be emphasized, contingent. Because the potentially confounding variables are unmeasured, one does not know their impact on the assignment process in the context of matching, or their relationship with the treatment variable or the outcome variable in the context of IV estimation. It follows, therefore, that the approaches discussed in this paper require some knowledge of the substantive process being studied, so that one can benchmark the strength of omitted potentially confounding variables

against observed variables. It seems reasonable to assume that many of the most powerful determinants of interesting outcome variables will indeed be measured, and therefore that one can with reasonable confidence guess at the potential impact of confounding variables. Whether these guesses are correct, of course, can only be tested via the gradual improvement in data collection, or via novel approaches such as the linking of experimental and nonexperimental data on some salient outcome (for example, Heckman et al. 1998). In the absence of such data, however, sensitivity analysis still provides an important tool for assessing the level of caution that one should use when interpreting the significance tests for causal effects that are produced with conventional estimators. To be able to formalize one's uncertainty in terms of the probable range of confounding effects from unmeasured variables and the likelihood that conventional estimates of causality are robust to such confounding effects is to deepen one's understanding about causal mechanisms.

## References

- Angrist, Joshua. 2003. *Treatment Effect Heterogeneity in Theory and Practice*. IZA (Institute for the Future of Labor) Discussion Paper No. 851: Bonn, Germany.
- Angrist, J., G.W. Imbens and D. Rubin, (1996). "Identification of Causal Effects Using Instrumental Variables," (with discussion) *Journal of the American Statistical Association* vol. 91, no 434, 444-472.
- Imbens, Guido W & Angrist, Joshua D, 1994. "[Identification and Estimation of Local Average Treatment Effects](#)," *Econometrica*, Vol. 62 (2) pp. 467-75.
- Heckman, James J. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources*. 32: 441-462.
- Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith. 2000. "The Economics and Econometrics of Active Labor Market Programs." Pg. 1865-2097 in *Handbook of Labor Economics, Vol. III*, edited by Orley Ashenfelter and David Card. Amsterdam: North Holland.
- Heckman, James J. and Edward J. Vytlacil. 2002. "Local Instrumental Variables." Unpublished paper. Chicago: Department of Economics.
- Holland, Paul. 1986. "Statistics and Causal Reasoning." *Journal of the American Statistical Association*. 81: 945-960.
- Lechner, M. 1999. "Nonparametric Bounds on Employment and Income Effects of Continuous Vocational Training in East Germany." *The Econometric Journal*, 2: 1-28.
- Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Outcomes." *American Economic Review*. 80: 319-323.
- Moffitt, Robert. 1996a. Selection Bias Adjustment in Treatment-Effect Models as a Method of Aggregation. NBER Technical Working Paper No. 187. Cambridge, MA: National Bureau of Economic Research.
- Moffitt, Robert. 1996b. "Identification of Causal Effects using Instrumental Variables: Comment." *Journal of the American Statistical Association*. 434: 462-465.
- Mortensen, Dale. 1986. "Job Search and Labor Market Analysis." Pg. 849-919 in *Handbook of Labor Economics, Vol. II*, edited by Orly Ashenfelter. Amsterdam: North-Holland.
- Rosenbaum, Paul R. 2002. *Observational Studies, Second Edition*. New York: Springer.
- Rosenbaum, Paul R., and Donald B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *The American Statistician* 39: 33-38.
- Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects." *The Annals of Statistics*. 6: 34-58.
- Rubin, Donald B. 1991. "Practical Implications of the Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism." *Biometrics* 47: 1213-1234.

- Sobel, Michael E. 1995. "Causal Inference in the Social and Behavioral Sciences." Pg. 1-38 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by Gerhard Arminger, Clifford C. Clogg, and Michael E. Sobel. New York: Plenum.
- Sobel, Michael E. 1996. "An Introduction to Causal Inference." *Sociological Methods and Research*. 24: 353-379.
- U.S. Bureau of the Census. 1991. *Survey of Income and Program Participation. User's Guide. Second Edition*. Washington, D.C.: U.S. Bureau of the Census.
- U.S. Department of Labor. 2002. *Comparison of State Unemployment Insurance Laws 2002*. Washington, D.C.: U.S. Department of Labor
- Vroman, W. 1990. "The Aggregate Performance of Unemployment Insurance, 1980-1985." Pg. 19-46 in *Unemployment Insurance*, edited by W.L. Hansen and J. F. Byers. Madison, WI: University of Wisconsin Press.
- Vroman, W. 2001. *Low Benefit Reciprocity in State Unemployment Programs*. The Urban Institute, mimeo.
- Winship, Christopher and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology*. 25: 659-706.

## Notes

<sup>1</sup> <http://www.wz-berlin.de/ars/ab/staff/gangl.en.htm>.

<sup>2</sup> These two quantities can differ because individuals who self-select themselves into treatment are typically not representative either of the entire population or of any subpopulation that can be defined with observable variables.

<sup>3</sup> The conditional ATE and ATT are still averages, because cases with identical  $X$  will not generally have identical values of  $U_0$  and  $U_1$ .

<sup>4</sup> That the matching estimator is an estimator in general for the ATT and not the ATE is easy to show. Imagine an experiment where the treatment is to press a button at which point an envelope is delivered to the subject via a mail slot. Imagine that the treatment is to be given the right to play a game with the assistant. Tickets to play the game are drawn at random. The price to play the game is \$10. Before one plays, the assistant holds a card that says that your reward for playing the game is either \$0 or \$20 with probability .5 for each outcome. If everyone given a ticket plays the game, the ATE is 0. If only those who know they will receive \$20 play the game, an averaging of actual rewards will give a consistent estimate of the ATT. The experimenter can only recover the ATE if she knows what the subjects know about their potential rewards from playing the game.

<sup>5</sup> These assumptions are (1) Stable unit treatment values (SUTVA), (2) random assignment to treatment, (3) the exclusion restriction (the outcome is independent of the IV, given the treatment status, and (4) nonzero causal effect of the IV on the treatment status, and (5) monotonicity (the probability of treatment is at least as high for a high value of the IV as for a low value for all units in the population).

<sup>6</sup> We have developed a Stata ado file that can implement these bounds in the case of propensity score matching for matched 1x1 pairs. The file is available at <http://www.wz-berlin.de/ars/ab/staff/gangl.en.htm>.

<sup>7</sup> Note that the Rosenbaum bounds bound the ATT, while the IV method bounds the LATE estimator. See Angrist (2002?) for the situation where the LATE might match the ATE even in the case of heterogeneous treatment effects.

<sup>8</sup> When the IV has multiple values and when the assumptions of the LATE are satisfied, then the IV sensitivity procedure must be applied at different points along the distribution of the IV in order to estimate the sensitivity of the LATE estimate at these different values of the IV.

---

<sup>9</sup> These data are described in U.S. Bureau of the Census (1991). The analysis sample is the inflow from employment into unemployment, and hence excludes spells of both first-time entrants to the labor force and (mostly) women returning to the labor market after family-related career interruptions.

<sup>10</sup> All wage data have been deflated to 1990 prices. In addition, the analyses exclude hourly real wage data below 1 and above 100 dollars (in 1990 prices), which is approximately equivalent to cutting off the top 1% of the observed wage distribution.

<sup>11</sup> The SUTVA must also be true for the IV estimator to yield the LATE.

<sup>12</sup> In most empirical applications, only a minority of cases is treated, and so matching extracts the subset of control cases which “best” matches with the treatment cases by some criterion. In cases where a majority of cases are treated, a matching procedure would be implemented by first randomly selecting a subsample of treated cases and then these cases would be matched with the control cases.

<sup>13</sup> This implies that we present linear probability models (LPM) for unemployment duration and regional mobility. We do so since the LPM yields parameter estimates on the probability scale that can be directly compared to the matching estimates, whereas in depending on scale, marginal effects calculated from logit or probit models would seem to offer only imperfect substitutes for this particular purpose.

<sup>14</sup> Strictly speaking, Rosenbaum’s procedure addresses hidden bias in general, i.e. comprises situations of both positive and negative self-selection into treatment status. Since the estimated treatment effect will actually represent a conservative effect estimate in a situation of negative self-selection, it is the case of differential assignment in accordance with outcomes (i.e. positive self-selection) that is of primary theoretical interest. Consequently, we restrict our attention to bounds for positive self-selection in the following.

<sup>15</sup> A “good cause” is a reason for quitting a job that does not disqualify one from receiving unemployment compensation. Our measure is constructed from U.S. Department of Labor (2002: 5.4ff.) and refers to the number of acceptable reasons for quitting. The measure reflects state legislation as of January 1, 2002. Apart from the fact that our main purpose is to illustrate the sensitivity analysis procedure, our substantive justification for the use of this measure would be to assume a certain degree of stability in state variation in UI generosity over time.

<sup>16</sup> We focus on disqualification policies as the appropriate policy dimension because we defined receipt of benefits as having had benefits during any (i.e. at least one) month in unemployment. In consequence, workers disqualified for failure to comply with search requirements at some point during an unemployment spell would still be counted as

---

having received benefits in our analyses. As we do not want to test for the effects of stringency of search requirements, but of having had access to benefits per se, we prefer a simpler and hopefully more robust time-constant measure, which at the same time permits a conservative hypothesis test in the sense that we assume benefit effects to follow even from partial benefit receipt during an unemployment spell.

<sup>17</sup> In supplementary analyses, we also used each of these components separately as the benchmark omitted variable. The results are similar, but the most conservative analysis is obtained when the two components are used simultaneously as the benchmark omitted variable.

<sup>18</sup> Note that if the correlation were negative instead of positive, the sign of  $\gamma$  would be reversed.

## Appendix A: The Rosenbaum Bounds Method

We recapitulate the Rosenbaum bounds method (Rosenbaum 2002) of sensitivity analysis for the estimation of treatment effects using data on matched pairs. Rosenbaum developed this approach to assess the impact of hidden bias on the computation of test statistics from the family of sign-score statistics, which are nonparametric tests that include Wilcoxon’s signed rank test and McNemar’s test. While Rosenbaum developed the theory for a more general case, we limit the discussion here to the case of matched pairs, which corresponds to the situation when propensity score analysis is used.

Test statistics in the family of sign score statistics have the form

$$T = t(Z, r) = \sum_{s=1}^S d_s \sum_{i=1}^2 c_{si} Z_{si} \quad (11)$$

where  $Z$  is the variable that registers which of each of the  $s$  pairs was treated, and  $r$  measures the outcome for each case in the  $S$  pairs.  $Z_{si}$  equals 1 if a case is treated, and 0 otherwise. “ $c$ ” is defined as follows:

$$\begin{aligned} c_{s1} = 1, c_{s2} = 0 & \text{ if } r_{s1} > r_{s2} \\ c_{s1} = 0, c_{s2} = 1 & \text{ if } r_{s1} < r_{s2} \\ c_{s1} = 0, c_{s2} = 1 & \text{ if } r_{s1} = r_{s2} \end{aligned}$$

Finally,  $d_s$  is the rank of  $|r_{s1} - r_{s2}|$  with average ranks used for ties. Essentially, the product of the  $c$  and  $Z$  variables cause pairs to be selected where the outcome for the treatment was greater than the outcome for the control. The ranks of these cases are summed and compared to the distribution of the test statistic under the null hypothesis that the treatment has no effect.

In the case where the assignment to treatment is not random, the above test statistic can be bounded. Rosenbaum proposes that one assume that there is an unmeasured variable  $u$  that

affects the probability of receiving the treatment. If we let  $\pi_i$  be the probability that the  $i$ th unit receives the treatment, and  $X$  are the observed covariates that determine treatment and that also determine the outcome variable, then the following treatment assignment equation applies

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \kappa(X_i) + \gamma U_i, \quad (12)$$

where  $0 \leq U_j \leq 1$

Rosenbaum shows that this relationship implies the following bounds on the ratio of the odds that either of two cases which are matched on  $X$  (or alternatively on the propensity score  $P(X)$ ) will receive the treatment

$$\frac{1}{\Gamma} \leq \frac{\pi_{s,1}(1-\pi_{s,2})}{\pi_{s,2}(1-\pi_{s,1})} \leq \Gamma \quad (13)$$

where  $s$  indexes the matched pair,  $s=1, \dots, S$ , and  $\Gamma = \exp(\gamma)$ .

Under the assumption that a confounding variable  $U$  exists, equation (11) becomes the sum of  $S$  independent random variables where the  $s$ th variations equals  $d_s$  with probability

$$p_s = \frac{c_{s1} \exp(\gamma u_{s1}) + c_{s2} \exp(\gamma u_{s2})}{\exp(\gamma u_{s1}) + \exp(\gamma u_{s2})}$$

and equals 0 with probability  $1 - p_s$ . Define

$$p_s^+ = \begin{cases} 0 & \text{if } c_{s1} = c_{s2} = 0 \\ \frac{\Gamma}{1+\Gamma} & \text{if } c_{s1} \neq c_{s2} \end{cases}$$

and

$$p_s^- = \begin{cases} 0 & \text{if } c_{s1} = c_{s2} = 0 \\ \frac{1}{1+\Gamma} & \text{if } c_{s1} \neq c_{s2} \end{cases}$$

Rosenbaum shows that for any specific  $\gamma$ , the null distribution of  $t(Z,r)$  is bounded by two known distributions for  $T^+$  and  $T^-$  where

$$E(T^+) = \sum_{s=1}^S d_s p_s^+$$

$$E(T^-) = \sum_{s=1}^S d_s p_s^-$$

$$Var(T^+) = \sum_{s=1}^S d_s^2 p_s^+ (1 - p_s^+)$$

$$Var(T^-) = \sum_{s=1}^S d_s^2 p_s^- (1 - p_s^-)$$

One can use these formulas to compute the significance level of the null hypothesis of no effect. For any specific  $\Gamma$ , one computes

$$(T - E(T^+)) / \sqrt{Var(T^+)}$$

and

$$(T - E(T^-)) / \sqrt{Var(T^-)}$$

where  $T$  is the Wilcoxon signed rank statistic. These two values give bounds of the significance level of a one-sided test for no effect of the treatment.

Under the assumption of an additive treatment effect, Rosenbaum (2002) also derives bounds on the Hodges-Lehmann point estimate of the treatment effect, enabling the researcher to frame the sensitivity analysis in the more common metric of an interval of point estimates rather than in terms of implied significance levels for the estimated treatment effect. To arrive at an interval of plausible point estimates given a specific bias level  $\Gamma$ , Rosenbaum defines the Hodges-Lehmann point estimate of the treatment effect

$$\hat{\delta} = \frac{\inf\{\delta: t' > t(Z, Y - \delta Z)\} + \sup\{\delta: t' < t(Z, Y - \delta Z)\}}{2}$$

Though not generally known, the expectation of that signed rank statistic is bounded by the expectations of  $T^+$  and  $T^-$  calculated at

$$t_{\min} = \frac{p^- S(S+1)}{2} \quad \text{and} \quad t_{\max} = \frac{p^+ S(S+1)}{2}$$

where

$$p^- = \frac{1}{1+\Gamma} \quad \text{and} \quad p^+ = \frac{\Gamma}{1+\Gamma}$$

as before. Since the bounds on the signed rank statistic are sharp, we can calculate an *interval* of point estimates consistent with these bounds by calculating the statistic at  $t = t_{\max}$  and  $t = t_{\min}$ . By similar reasoning, Rosenbaum also derives bounds for the confidence interval of the point estimate.

## Appendix B – Stratified nearest neighbor matching

1. We estimate a logistic regression for the observed treatment variable  $D$  (UI benefit status). The covariate vector  $X$  of the model includes a set of standard covariates that are thought to potentially affect both  $D$  and the outcome of interest  $Y$ .

2. Based on the estimated logistic model, we compute the propensity score as the predicted

probability of receiving benefits from 
$$\frac{\exp(\hat{b}'X)}{1 + \exp(\hat{b}'X)}$$
. We also calculate the

variance of the propensity score, and choose a caliper  $w$  that reflects the degree of similarity (measured in standard deviation units) required to form matches of treatment and control observations.

3. From the estimated propensity score, we form 1x1 matches of treatment and control cases by stratified random-order nearest-neighbor matching as follows:

(a) Within state and time points defining the strata, randomly select a person  $i$  from the treatment sample of workers with unemployment benefits.

(b) Find all observations  $j$  in the control sample of workers without unemployment benefits in that state at that particular time point that satisfy the condition

$$\hat{b}'Z_j \in \hat{b}'Z_i \pm w\sqrt{\text{var}(\hat{b}'Z_i)}, \text{ i.e. whose propensity scores are sufficiently}$$

similar to  $i$ , with the degree of similarity being determined by the width of the confidence interval around  $\hat{b}'Z_i$ .

(c) If there is no observation  $j$  falling in the required interval, remove the treatment case  $i$  from the sample. There is no adequate control case

available in the sample at the required level of observational similarity.

- (d) If there is one or more observations  $j$  falling in the required interval, form a matched pair of cases between  $i$  and the observation  $j$  that exhibits the smallest propensity score distance  $|\hat{b}'Z_i - \hat{b}'Z_j|$ . (I.e. we refrain from matching on additional covariates not included in  $Z$  in the following.)  
Remove the matched pair from the sample of observations (i.e. we perform matching without replacement of control cases).
- (e) Repeat steps (a)-(d) until no treated cases are left for matching within the particular stratum, then repeat steps (a)-(d) for all strata defined by U.S. states times four three-year time periods (1984-86, 1987-89, 1990-92, 1993-95).

## Appendix C – Caliper size, sample sizes, overt and hidden bias in wage change models

Caliper $w$	N matches (% matched)	Bias   on propensity score (% reduction)	Treatment effect $\bar{\delta}$ (s.e.) <sup>1)</sup>	Critical level of $\Gamma$ (hidden bias)
Standard nearest-neighbor matching				
0.500	5,040 (93.8%)	46.0 (45.2%)	+0.049 (.039)	1.05-1.10
0.250	4,668 (86.8%)	35.8 (57.3%)	+0.087** (.013)	1.15-1.20
0.100	4,589 (85.4%)	33.4 (60.2%)	+0.084** (.015)	1.10-1.15
0.050	4,451 (82.8%)	29.4 (64.9%)	+0.079** (.018)	1.10-1.15
0.010	3,900 (72.6%)	12.5 (85.1%)	+0.071** (.024)	1.10-1.15
0.001	2,841 (52.9%)	24.2 (71.1%)	+0.028 (.021)	1.00
Nearest-neighbor matching stratified by state and time period				
0.500	5,008 (93.2%)	30.4 (63.7%)	+0.033 (.020)	1.00-1.05
0.250	4,116 (76.6%)	13.1 (84.3%)	+0.094** (.022)	1.15-1.20
0.100	3,700 (68.8%)	11.6 (86.2%)	+0.086** (.029)	1.15-1.20
0.050	3,263 (60.7%)	6.7 (92.0%)	+0.072** (.025)	1.10-1.15
0.010	2,066 (38.4%)	1.2 (98.6%)	+0.055** (.027)	1.05-1.10
0.001	618 (11.5%)	0.1 (99.9%)	+0.016 (.049)	1.00

Notes: <sup>1</sup> Bootstrap standard errors, N=100 replication samples. Statistical significance levels at \*\* p<.01 and \* p<.05, respectively.

Source: Survey of Income and Program Participation, Panels 1984, 1986, 1988, 1990, 1992 and 1993.

## Appendix D – Bivariate regressions results for group-level covariates

	Log wage change	Prob of Unemployment Exit T<3mo.	Regional mobility	$r(\bar{X}, \bar{D})$
Years of education	-0.003 (.024)	0.070* (.031)	0.075** (.025)	0.035
Unemployment rate	-0.001 (.004)	-0.021** (.005)	-0.013** (.004)	0.227

Notes: Results for regressions of state-level mean outcomes on the state-level mean covariate (standard errors in parentheses; significance levels at \*\* p<.01 and \* p<.05); last column gives aggregate correlation between state-level covariate and state-level UI coverage rates.

Source: Survey of Income and Program Participation, Panels 1984, 1986, 1988, 1990, 1992 and 1993.

**Table 1 – Logistic Regression Model for Receiving Unemployment Benefits, given Unemployment**

	Coefficient	Standard Error
Log previous wage	1.927**	(.093)
Log previous wage squared	-0.288**	(.017)
Labor force experience (years)	0.092**	(.004)
Labor force experience squared (*100)	-0.149**	(.009)
Tenure with previous employer (months)	0.007**	(.001)
Tenure with previous employer squared (*100)	-0.002**	(2.0e <sup>-4</sup> )
Education (Reference Category: less than High School)		
- High School	0.197**	(.039)
- Some College	0.112**	(.044)
- B.A. / Associate degree	-0.033	(.059)
- M.A. and above	-0.235**	(.082)
Female	-0.164**	(.029)
Race (Reference Category: White)		
- Black	-0.392**	(.043)
- Hispanic	0.131**	(.046)
- Other Non-White	-0.007	(.078)
Constant	-4.140**	(.120)
N (unemployment spells)	26,284	
Log-likelihood	-15,053.7	
Pseudo-R <sup>2</sup>	0.114	

Source: Survey of Income and Program Participation, Panels 1984, 1986, 1988, 1990, 1992 and 1993.

**Table 2 - Propensity Score Matching and Covariate Balance**

Variable	Sample	$\bar{X}_{Treated}$	$\bar{X}_{Controls}$	bias (%)	% reduction in  bias
<i>Wage change model</i>					
<i>(N=3,263 matched pairs)</i>					
Propensity score	Unmatched	0.444	0.307	83.88	
	Matched	0.361	0.370	6.68	92.0
Ln previous wage	Unmatched	2.417	2.021	51.21	
	Matched	2.204	2.242	4.98	90.3
Labor force experience (years)	Unmatched	16.18	9.76	60.03	
	Matched	12.37	12.55	1.78	97.0
Tenure with previous employer (months)	Unmatched	31.51	12.67	36.00	
	Matched	15.36	13.73	4.24	88.2
Years of education	Unmatched	12.56	12.49	3.73	
	Matched	12.56	12.67	5.73	-53.5
Female	Unmatched	0.346	0.444	19.84	
	Matched	0.412	0.409	0.50	97.5
Black	Unmatched	0.093	0.138	14.11	
	Matched	0.123	0.106	5.30	62.5
Hispanic	Unmatched	0.117	0.106	3.63	
	Matched	0.148	0.132	4.42	-21.8
Other non-white	Unmatched	0.031	0.031	0.20	
	Matched	0.037	0.037	0.49	-148.7
<i>Unemployment duration model</i>					
<i>(N=5,550 matched pairs)</i>					
Propensity score	Unmatched	0.442	0.301	84.21	
	Matched	0.362	0.370	5.75	93.2
<i>Regional mobility model</i>					
<i>(N=1,818 matched pairs)</i>					
Propensity score	Unmatched	0.457	0.317	83.47	
	Matched	0.355	0.364	6.75	91.9

Source: Survey of Income and Program Participation, Panels 1984, 1986, 1988, 1990, 1992 and 1993.

**Table 3 – Treatment Effects of Unemployment Benefit Receipt, Matching Estimates**

	$Y_T$	Raw $Y_C$	Matched $Y_C$	$\bar{\delta} = \Delta Y$ <sup>1</sup>	$\beta_{OLS}$ <sup>2</sup>
Ln wage change <sup>3</sup> (N=3,263 matched pairs)	-0.044 (0.957)	-0.036 (0.964)	-0.116 (0.891)	+0.072** (.025)	+0.073** (.012)
Pr(exit   T≤3 months) (N=5,550 matched pairs)	0.459	0.667	0.666	-0.207** (.014)	-0.208** (.007)
Pr(regional mobility) (N=1,818 matched pairs)	0.029	0.061	0.056	-0.027** (.010)	-0.033** (.005)

Notes: <sup>1</sup> Bootstrap standard errors, N=100 replication samples.

<sup>2</sup> Heteroskedasticity-consistent standard errors in parentheses.

<sup>3</sup> Exponentiated change scores in parentheses.

Statistical significance levels at \*\* p<.01 and \* p<.05, respectively.

Source: Survey of Income and Program Participation, Panels 1984, 1986, 1988, 1990, 1992 and 1993.

**Table 4 – Rosenbaum Bounds for Unemployment Benefit Treatment Effects**

	$\Gamma$	p-critical <sup>1</sup>	Hidden bias equivalents <sup>2</sup>		
			Ln prev. wage	Experi-ence	Tenure
<i>Wage change model</i> (N=3,263 matched pairs)	1.00	<.0001	0	0	0
	1.05	0.001	0.077	0.90	8.06
	1.10	0.028	0.150	1.79	16.17
	1.15	0.206	0.228	2.69	24.36
	1.20	0.590	0.309	3.60	32.67
	1.25	0.892	0.395	4.55	41.44
<i>Unemployment duration model</i> (N=5,550 matched pairs)	1.00	<.0001	0	0	0
	1.50	<.0001	1.177	9.86	93.25
	2.00	<.0001	-	-	-
	2.10	0.001	-	-	-
	2.20	0.028	-	-	-
	2.30	0.203	-	-	-
	2.40	0.580	-	-	-
	2.50	0.883	-	-	-
<i>Regional mobility model</i> (N=1,818 matched pairs)	1.00	<.0001	0	0	0
	1.20	0.001	0.309	3.60	32.67
	1.40	0.016	0.711	7.55	69.87
	1.50	0.041	1.177	9.86	93.25
	1.60	0.085	-	12.79	180.37
	1.80	0.239	-	-	-
	2.00	0.453	-	-	-

Notes: <sup>1</sup>p-critical is p<sup>+</sup> for wage change model, p<sup>-</sup> for unemployment duration and regional mobility models.

<sup>2</sup>Hidden bias equivalents are computed at the empirical mean of covariates. Due to the nonlinear specification of covariate effects, equivalents cannot be computed at all levels of  $\Gamma$ .

Source: Survey of Income and Program Participation, Panels 1984, 1986, 1988, 1990, 1992 and 1993.

**Table 5 – IV Estimates of Unemployment Benefit Effects**

	Benefit status instrumented with		
	High vs. low generosity states (5 levels)	State dummies (45 levels)	State policy Z: Liberal good cause regulation
<i>Log wage change</i>	0.355** (.087)	0.123* (.072)	0.432** (.159)
<i>Unemployment duration</i> <i>Pr(exit T&lt;3)</i>	-0.475** (.057)	-0.417** (.051)	-0.345** (.100)
<i>Regional mobility</i> <i>Pr(state change T&lt;18)</i>	-0.116** (.037)	-0.125** (.031)	-0.222** (.064)

Notes: Models include the same set of covariates as the one for estimating the propensity score (cf. Table 1 above).  
Source: Survey of Income and Program Participation, Panels 1984, 1986, 1988, 1990, 1992 and 1993.

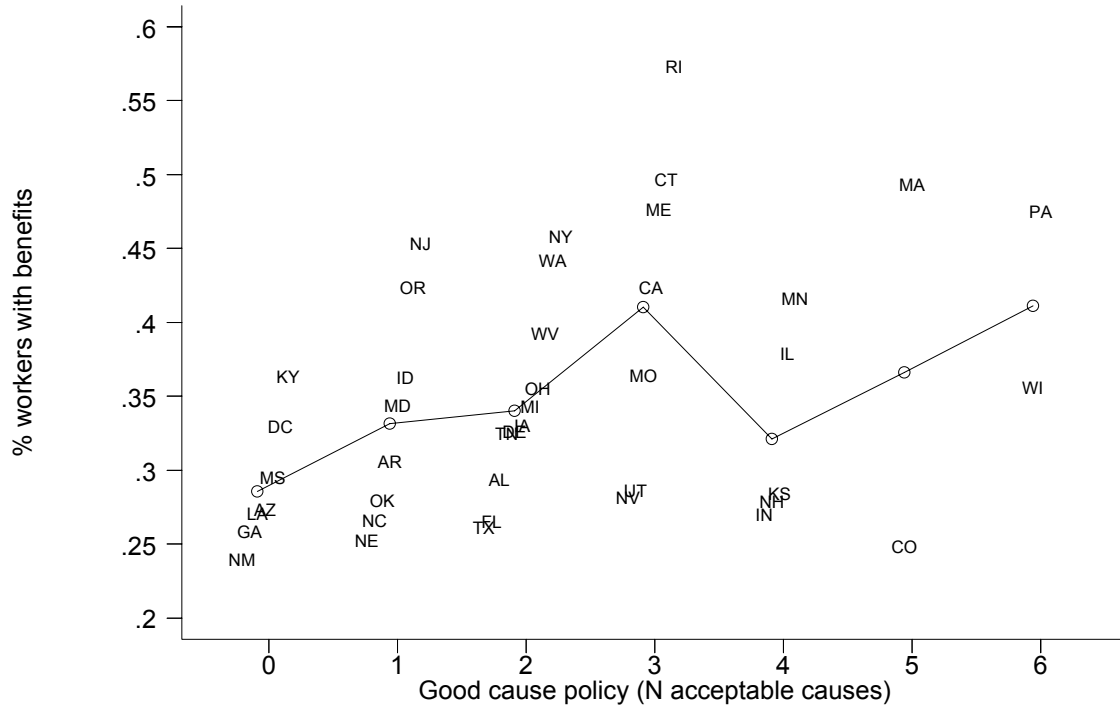
**Table 6 – Sensitivity Analysis for IV Estimates of Unemployment Benefit Effects**

Instrument	High vs. low generosity states (5 levels)				State policy Z: Liberal good cause regulation			
	Corr(W,D)				Corr(W,D)			
	.25	.50	.80	.95	.25	.50	.80	.95
$\gamma(W)$	<i>Log wage change</i>							
+0.01	<.001	0.001	0.006	0.013	0.013	0.038	0.113	0.175
+0.02	<.001	0.001	0.006	0.013	0.038	0.200	0.607	0.796
+0.03	<.001	0.005	0.040	0.093	0.096	0.534	0.960	0.995
+0.05	0.002	0.049	0.409	0.684	0.353	0.974	1.000	1.000
+0.07	0.009	0.241	0.901	0.989	0.708	0.974	1.000	1.000
+0.10	0.049	0.763	1.000	1.000	0.974	1.000	1.000	1.000
$\gamma(W)$	<i>Unemployment duration</i>							
-0.01	<.001	<.001	<.001	<.001	<.001	0.030	0.169	0.309
-0.02	<.001	<.001	<.001	<.001	0.030	0.365	0.933	0.992
-0.03	<.001	<.001	<.001	<.001	0.133	0.883	1.000	1.000
-0.05	<.001	<.001	0.004	0.056	0.664	1.000	1.000	1.000
-0.07	<.001	<.001	0.350	0.864	0.975	1.000	1.000	1.000
-0.10	<.001	0.109	0.999	1.000	1.000	1.000	1.000	1.000
$\gamma(W)$	<i>Regional mobility</i>							
-0.01	0.004	0.015	0.062	0.109	0.005	0.040	0.237	0.420
-0.02	0.015	0.130	0.544	0.767	0.040	0.488	0.979	0.999
-0.03	0.050	0.462	0.961	0.996	0.187	0.954	1.000	1.000
-0.05	0.270	0.975	1.000	1.000	0.796	1.000	1.000	1.000
-0.07	0.663	1.000	1.000	1.000	0.994	1.000	1.000	1.000
-0.10	0.975	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Notes: Statistical significance levels for one-tailed tests.

Source: Survey of Income and Program Participation, Panels 1984, 1986, 1988, 1990, 1992 and 1993.

**Figure 1 – State variation in access to unemployment benefits, by good cause policy**



Notes: 1984-1995 average data; data unadjusted for differences in the structure of work forces between states.  
 Source: Survey of Income and Program Participation, Panels 1984, 1986, 1988, 1990, 1992 and 1993.