

Comparing Experimental and Nonexperimental Methods: What Lessons Have We Learned Four Decades after LaLonde (1986)?

Guido W. Imbens and Yiqing Xu

In 1986, Robert LaLonde published a paper based on part of his PhD thesis (LaLonde 1986), which has had profound impact on both the methodological and empirical literatures on estimating causal effects. As of May 2025, this paper has been cited over 3,000 times, a number that only partially reflects its tremendous influence on the field of causal inference and the credibility revolution (Angrist and Pischke 2010).

The context for LaLonde's (1986) paper was the National Supported Work demonstration program. This program targeted individuals with extremely poor employment prospects: for females, recipients of Aid to Families with Dependent Children; for males, ex-drug addicts, ex-criminal offenders, and high school drop-outs. Although about 10,000 people participated across 15 cities, the evaluation focused on approximately 6,600 participants in 10 cities. Approximately half were randomly assigned to a control group, while the other half received up to twelve months of training in small groups, with a supervisor and a counselor, followed by efforts to place them with outside employers. According to the Manpower Development Research Corporation (MDRC), which operated the program and conducted

■ *Guido W. Imbens is an Applied Econometrics Professor and a Professor of Economics at the Graduate School of Business and the Department of Economics, Stanford University, Stanford, California. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Yiqing Xu is an Assistant Professor in Political Science at the Department of Political Science and a W. Glenn Campbell and Rita Ricardo-Campbell National Fellow at the Hoover Institution at Stanford University, Stanford, California. Their email addresses are imbens@stanford.edu and yiqingxu@stanford.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at <https://doi.org/10.1257/jep.20251440>.

the evaluation, the National Supported Work program substantially increased 1978 earnings for female participants, particularly those without prior job experience; in contrast, the effects for male participants were smaller and, for some subgroups, essentially nonexistent (MDRC 1980).

LaLonde (1986) used this experimental evaluation to address a broader methodological question: whether the state-of-the-art nonexperimental evaluation methods of that time could replicate experimental benchmarks—that is, estimates obtained from randomized controlled trials. By nonexperimental methods, we mean statistical approaches that estimate a program’s impact using naturally occurring variation in treatment assignment, without randomization or experimental control. LaLonde’s conclusion was sharply negative regarding the credibility of the nonexperimental methods he examined. The abstract (p. 604) noted: “This comparison shows that many of the econometric procedures do not replicate the experimentally determined results, and it suggests that researchers should be aware of the potential for specification errors in other nonexperimental evaluations.” He concluded (p. 617): “Policymakers should be aware that the available nonexperimental evaluations of employment and training programs may contain large and unknown biases resulting from specification errors.”

Around the time LaLonde’s paper was published, several other studies made similar points about the credibility of nonexperimental methods, including work by LaLonde’s thesis advisers Ashenfelter and Card (1985), as well as Fraker and Maynard (1987) and Heckman, Hotz, and Dabos (1987). One reason LaLonde’s study became so influential is that the full data underlying the original analysis later became publicly available. As part of a research project initiated in a 1996 graduate class taught by Imbens and Rubin at Harvard, Dehejia and Wahba (1999a, 2002) retrieved LaLonde’s original male subsample data—stored on tapes—by locating a tape reader capable of reading the files. The data are now publicly available at Dehejia and Wahba (1999b). We refer to a subsample of these data that is most commonly used in the literature as the LaLonde-Dehejia-Wahba (LDW) data. More recently, Calónico and Smith (2017) also reconstructed the female samples. These public datasets are highly valuable for both teaching and future research and are provided alongside this paper in the Supplemental Appendix and Imbens and Xu (2025), with code to replicate the estimates reported here.

The methodological literature on nonexperimental evaluation has advanced significantly. Nearly four decades after LaLonde’s paper, the answer to his original question—whether nonexperimental methods can successfully replicate experimental benchmarks—is more nuanced than his initial conclusion: Sometimes they can, and we now have better tools to help assess when they are likely to succeed.

Here are five key lessons that have emerged from this literature since the publication of LaLonde (1986). First, the unconfoundedness assumption has become central to modern methods for nonexperimental data. Unconfoundedness means that, conditional on pretreatment variables or covariates, treatment is as if randomly assigned. This assumption elegantly separates the underlying

identification assumptions from functional form considerations. It emphasizes design—how treatment is assigned—over full specification of the data-generating process.

Second, inspecting and ensuring overlap is critical in the process of estimating treatment effects. Overlap means that the covariate distributions in treatment and control groups have common support. Improved overlap reduces sensitivity to the choice of estimation strategy and improves the robustness of estimates.

Third, the use of propensity scores, defined as the probability of receiving treatment given observed covariates, has become widespread. The propensity score, introduced in Rosenbaum and Rubin (1983a), had only just entered the economics literature when LaLonde (1986) was writing his thesis. Since then, propensity scores have become a core component of many analyses that rely on unconfoundedness, including for assessing overlap, but also for estimation.

Fourth, researchers have increasingly taken heterogeneity in treatment effects seriously. This includes both methods for estimating average treatment effects in the presence of heterogeneity, but also going beyond average treatment effects to examine estimates of treatment effect heterogeneity, such as conditional treatment effects and quantile treatment effects (for example, Wager and Athey 2018). These approaches provide a deeper understanding of how the treatment affects different individuals or segments of the outcome distribution.

Finally, validation exercises—particularly placebo tests—are essential for assessing key assumptions and evaluating the credibility of causal claims. In a placebo test, researchers use an outcome that should not be affected by the treatment. For example, LaLonde (1986) used 1975 earnings as a placebo outcome, because it predates the training program. A non-zero estimated effect on the placebo outcome indicates potential unobserved confounding. Placebo tests have been important diagnostic tools in modern nonexperimental research since the credibility revolution.¹

To illustrate these lessons, we reexamine the original and reconstructed LaLonde datasets. We show that, once overlap is ensured, various modern methods yield similar estimates. However, these estimates lack a causal interpretation if unconfoundedness is violated. Validation exercises such as placebo tests are critical for assessing the credibility of unconfoundedness. For the LaLonde datasets, placebo estimates generally do not support the unconfoundedness assumption.

¹ We focus on five issues specific to the setting studied in LaLonde (1986): the evaluation of an individual-level intervention using detailed background information on participants. While the broader literature on causal inference has expanded substantially over the past four decades (for an illustration of the changes in methodologies used over this period, see Currie, Kleven, and Zwiars 2020), we do not attempt to cover that broader landscape here. For more comprehensive overviews, see surveys such as Imbens and Wooldridge (2009) and Abadie and Cattaneo (2018), as well as a number of textbooks (Angrist and Pischke 2008; Imbens and Rubin 2015; Cunningham 2021; Huntington-Klein 2021; Huber 2023; Ding 2024; Wager 2024; Chernozhukov et al. 2024).

LaLonde's Data and Findings

We begin by describing the data used in LaLonde (1986) and outlining the main econometric approaches and results presented in the paper. We then introduce the LaLonde-Dehejia-Wahba data and examine some of the contemporary methodological responses to LaLonde's findings.

LaLonde's Data

Although the evaluation by the Manpower Development Research Corporation included 6,600 individuals, the experimental samples LaLonde used to evaluate nonexperimental methods were much smaller. For males, the sample included 722 individuals (297 treated and 425 controls), and for females, 1,158 individuals (600 treated and 585 controls). This substantial reduction in sample size stemmed from three main factors. First, many participants failed to complete follow-up interviews needed to collect post-program earnings data. Second, due to budget constraints, the MDRC team randomly selected subsamples for the 27- and 36-month follow-up interviews. Third, LaLonde excluded male participants who entered the program before January 1976 or were still enrolled in January 1978.

To conduct a nonexperimental analysis of the effects of the training program, LaLonde combined the experimental treatment group with comparison groups drawn from external, nonexperimental datasets used as controls. For both female and male participants, he used two main sources. The first, CPS-SSA-1, is based on Westat's Matched Current Population Survey–Social Security Administration File and includes all females or males under age 55 who met Westat's eligibility criteria. The second, PSID-1, comes from the Panel Study of Income Dynamics and includes all female or male household heads under age 55 from 1975 to 1978 for males and 1979 for females, excluding those who identified as retired in 1975. The age cutoff was intended to improve comparability between the comparison group and the experimental sample. In Table 1, columns 1–4 present summary statistics for the male samples in the experimental data and the comparison groups LaLonde used. Respondents in both the CPS and PSID comparison groups were, on average, older, more educated, substantially less likely to be high school dropouts, more likely to be married, and less likely to be Black or Hispanic than participants in the National Supported Work experiment. They also had much higher earnings in the years before the program. For example, average 1975 earnings were \$3,000 in the experimental data, compared to \$14,000 in the CPS-SSA-1 sample and \$19,000 in the PSID-1 sample. Notably, the comparison groups are much larger: the male CPS-SSA-1 sample includes 15,922 observations, and the male PSID-1 sample contains 2,490 observations. To improve comparability with the experimental sample, LaLonde dropped from these datasets individuals based on employment status, time of survey, and poverty status, creating four additional, smaller comparison groups.

Columns 5 and 6 of Table 1 present summary statistics for the experimental treated and control units in the LaLonde-Dehejia-Wahba data (discussed further

Table 1

Descriptive Statistics: LaLonde and LaLonde-Dehejia-Wahba Male Samples

	<i>LaLonde experimental</i>		<i>LaLonde comparison groups</i>		<i>LaLonde-Dehejia- Wahba experimental</i>	
	<i>Treated</i> (1)	<i>Control</i> (2)	<i>CPS-SSA-1</i> (3)	<i>PSID-1</i> (4)	<i>Treated</i> (5)	<i>Control</i> (6)
Age	24.63 (6.69)	24.45 (6.59)	33.23 (11.05)	34.85 (10.44)	25.82 (7.16)	25.05 (7.06)
Years of school	10.38 (1.82)	10.19 (1.62)	12.03 (2.87)	12.12 (3.08)	10.35 (2.01)	10.09 (1.61)
Proportion high school dropouts	0.73 (0.44)	0.81 (0.39)	0.30 (0.46)	0.31 (0.46)	0.71 (0.46)	0.83 (0.37)
Proportion married	0.17 (0.37)	0.16 (0.36)	0.71 (0.45)	0.87 (0.34)	0.19 (0.39)	0.15 (0.36)
Proportion Black	0.80 (0.40)	0.80 (0.40)	0.07 (0.26)	0.25 (0.43)	0.84 (0.36)	0.83 (0.38)
Proportion Hispanic	0.09 (0.29)	0.11 (0.32)	0.07 (0.26)	0.03 (0.18)	0.06 (0.24)	0.11 (0.31)
Real earnings in 1975 (thousand)	3.07 (4.87)	3.03 (5.20)	13.65 (9.27)	19.06 (13.60)	1.53 (3.22)	1.27 (3.10)
Proportion unemployed in 1975	0.37 (0.48)	0.42 (0.49)	0.11 (0.31)	0.10 (0.30)	0.60 (0.49)	0.68 (0.47)
Real earnings in 1974 (thousand)	NA	NA	14.02 (9.57)	19.43 (13.41)	2.10 (4.89)	2.11 (5.69)
Proportion unemployed in 1974	NA	NA	0.12 (0.32)	0.09 (0.28)	0.71 (0.46)	0.75 (0.43)
Observations	297	425	15,922	2,490	185	260

Note: Standard deviations are shown in parentheses. The experimental data were drawn from the National Supported Work program. Tables 3, 5, and 6 in LaLonde (1986) use the data summarized in columns 1–4, while Dehejia and Wahba (1999a) primarily rely on the data in columns 3–6.

below), a subset of the original LaLonde male sample selected for containing information on 1974 earnings. Our reanalysis focuses on this dataset because it is widely used in the post-LaLonde methodological literature and includes two pretreatment outcomes—earnings in 1974 and 1975—which enable adjustment for longer earnings histories and placebo analyses. Results for the original LaLonde male samples and reconstructed female samples are reported in the Supplemental Appendix.

Econometric Approaches in LaLonde (1986)

To estimate the causal effect of the National Supported Work program on 1978 earnings using both experimental and nonexperimental data, LaLonde employed a variety of models that fall into two broad categories: regression methods, in which earnings serve as the outcome (referred to as the “earnings equation”), and selection models that also include a “participation equation,” where the outcome is program participation. In LaLonde’s paper, Tables 4 and 5 present the regression

results, while Table 6 reports the selection model results. These tables are reproduced in the Supplemental Appendix.

LaLonde reported training effect estimates for both female and male participants using seven estimators across six comparison groups: two large groups described in Table 1 and the aforementioned four subgroups defined through additional selection criteria to improve overlap with the National Supported Work demonstration sample. All regression models were linear and implicitly assumed constant treatment effects. Of the seven models, two are simple regressions estimated with or without controls for age, education, and race (but notably excluding 1975 earnings). Two models use a difference-in-differences estimator, replacing the outcome with the change in earnings between 1975 and 1978, estimated with and without controlling for age. Two more use a quasi-difference-in-differences specification, including 1975 earnings on the right-hand side to account for transitory shocks—known as the “Ashenfelter dip” (Ashenfelter 1978)—again estimated with and without controlling for age. The final specification includes all available pretreatment covariates, including 1975 earnings, 1975 unemployment status, and marital status.

In an additional analysis in the spirit of the modern causal literature, LaLonde also conducted placebo tests using 1975 earnings as the outcome. Because the training occurred after 1975, the true treatment effect on this placebo outcome is zero. Indeed, the placebo estimates from the linear regression approach on the experimental sample were close to zero.

LaLonde emphasized several key findings from the linear regression results. First, using the experimental data, all seven estimators produced similar estimates: around \$851 for female participants and \$886 for male participants. Second, when using nonexperimental comparison groups, the estimates diverged sharply from these benchmarks, often yielding large, negative values in both female and male samples, with modest standard errors. Third, the estimates from nonexperimental data varied widely across specifications, and goodness-of-fit tests offered little guidance for selecting models that aligned with the experimental results. Taken together, these findings led LaLonde to conclude that the regression adjustment methods commonly used at the time were not credible when applied to nonexperimental data.

In addition to the regression models, which rely on exogeneity of the treatment indicator, LaLonde also presented results from selection models that allow for potential endogeneity. These models use the two-step estimator proposed by Heckman (1978), which permits correlation between the error terms in the earnings and participation equations. Identification in this framework relies either on exclusion restrictions—variables that appear in the participation equation but not in the earnings equation—or on functional form and distributional assumptions (linearity and joint normality of the error terms). For both female and male samples, LaLonde used three comparison groups (the experimental controls, CPS-SSA-1, and PSID-1) and estimated four specifications for females and three for males. Each specification used a different set of excluded variables. A priori, no single specification was clearly more defensible based on economic or econometric reasoning.

LaLonde found that the selection model estimates using the experimental data remained close to \$851 for females and \$886 for males. However, those using nonexperimental data again varied substantially and deviated from the experimental benchmarks. He concluded that, although the two-step procedure brought the estimates somewhat closer to the benchmarks, it still yielded a “considerable range of imprecise estimates” (LaLonde 1986, p. 617).

The LaLonde-Dehejia-Wahba Data

The data reconstructed by Dehejia and Wahba (1999a) focused exclusively on male participants, noting that “estimates for this group were the most sensitive to functional-form specification” (p. 1054). They constructed a subsample from LaLonde’s original data consisting of individuals with available information on 1974 earnings and unemployment status. They argued that this subsample remained a valid experimental sample because it was constructed solely using pretreatment information—such as month of assignment and employment history—thereby preserving the orthogonality of treatment assignment with respect to observed and unobserved characteristics. Notably, this subsample includes only 62 percent of the original treatment group in LaLonde’s study. For nonexperimental controls, they used subsets of the same datasets LaLonde employed, restricted to units with 1974 earnings and unemployment data. This collection, now widely known as the LaLonde-Dehejia-Wahba (LDW) data, has become a standard benchmark in the causal inference literature. In particular, most methodological studies focus on the version combining the experimental treated units with CPS-SSA-1 controls. For brevity, we refer to the LaLonde-Dehejia-Wahba experimental sample as LDW-Experimental, and the samples combining these treated units with CPS-SSA-1 and PSID-1 controls as LDW-CPS and LDW-PSID, respectively.

As noted earlier, columns 5 and 6 of Table 1 report summary statistics for the LDW-Experimental sample. Compared to LaLonde’s original male sample, participants in this subsample had higher unemployment rates and lower average earnings in 1975. The inclusion of 1974 earnings—available only in the LaLonde-Dehejia-Wahba data—further suggests that many participants faced long-term unemployment. These differences may help explain why the estimated training effect in this sample (\$1,794) is more than twice that of LaLonde’s original male sample (\$886).

Subsequent Literature

The publication of LaLonde (1986) sparked a debate in the applied econometrics literature. For example, Heckman and Hotz (1989) responded to LaLonde’s critique of nonexperimental evaluation methods by advocating the use of specification tests to rule out particularly poor estimators. However, this approach did not offer a clear way to distinguish among the many estimators that fit the data reasonably well but rely on different identification assumptions. As a result, it gained limited traction in subsequent work.

In contrast, Dehejia and Wahba (1999a) proposed alternative, more flexible estimators to address LaLonde’s challenge. Their proposals included approaches

based on propensity score stratification and matching. Their estimates closely matched the experimental benchmark, leading them to conclude (p. 1062), in sharp contrast to LaLonde: “[T]he estimates of the training effect for LaLonde’s . . . dataset are close to the benchmark experimental estimates and are robust to the specification of the comparison group and to the functional form used to estimate the propensity score. Our methods succeed for a transparent reason: They use only the subset of the comparison group that is comparable to the treatment group, and discard the complement.”

The contrast between the conclusions of Dehejia and Wahba (1999a) and LaLonde (1986) led to a wave of methodological research aimed at probing and reconciling these findings. We turn to these developments in the next section.

Methodological Improvements since LaLonde (1986)

Over the past four decades, the applied econometrics literature has made significant progress—most notably in clarifying estimands (what is being estimated) and the assumptions required for identification (how to use data to recover them). Once these assumptions are clearly stated, researchers can assess how plausible they are. To facilitate discussion of methodological advances since LaLonde (1986), we begin with the “potential outcomes framework,” then introduce the main causal estimand and a closely related statistical estimand. We then outline the key assumptions required for nonexperimental data to identify the causal estimand.

Causal and Statistical Estimands

In our discussion, we begin with the potential outcome framework, originally developed by Jerzy Sława-Neyman (1990) in the context of randomized experiments and extended to nonexperimental settings by Donald Rubin (Rubin 1974, 2006; Imbens and Rubin 2015). For each individual i , two potential outcomes are postulated: $Y_i(0)$ and $Y_i(1)$. In the LaLonde setting, they represent the individual’s 1978 earnings had the individual not participated in the program and had the person participated, respectively. The causal effect for individual i is the difference $Y_i(1) - Y_i(0)$. Let W_i denote the binary treatment indicator, equal to 1 if the individual participated in the program and 0 otherwise. The observed outcome is linked to the potential outcomes by the identity $Y_i = (1 - W_i)Y_i(0) + W_iY_i(1)$. We also observe pretreatment characteristics or covariates X_i . In the LaLonde study, these include age, years of schooling, high school dropout status, marital status, and indicators for Black and Hispanic backgrounds.² Following Dehejia and Wahba (1999a), we may augment this vector to include 1974 and 1975 earnings and indicators for zero earnings (unemployment) in each year.

²The modern methodological literature has paid particular attention to settings where the vector of pretreatment variables is high-dimensional.

Our primary goal is to estimate the average treatment effect on the treated (ATT), which represents the individual-level causal effect averaged over for those who actually received the treatment.³

$$ATT = E[Y_i(1) - Y_i(0) | W_i = 1] = E[Y_i | W_i = 1] - E[Y_i(0) | W_i = 1].$$

In LaLonde's context, it means the average treatment effect of the job training program on individuals who actually participated. Because $Y_i(1) = Y_i$ for treated individuals (individuals with $W_i = 1$), the second equality follows. The first term on the right-hand side, $E[Y_i | W_i = 1]$, can be directly estimated from the observed data. The second term, $E[Y_i(0) | W_i = 1]$, cannot, because it involves counterfactual outcomes—in this case, what participants' earnings would have been had they not participated in the program.

Most analyses of the LaLonde data that explicitly allow for heterogeneous treatment effects focus on the ATT, as it makes little sense to estimate or even contemplate the effect of the program for nonparticipants with stable jobs and high earnings. In other contexts, researchers may be interested in the average treatment effect (ATE), which captures the average effect across the entire population, while the ATT focuses specifically on the treated group. LaLonde (1986) did not explicitly distinguish between different estimands, as his analysis did not consider treatment effect heterogeneity.

Because we cannot observe the counterfactual outcomes, we cannot directly estimate the ATT. This is the core of what Holland (1986) termed the “fundamental problem of causal inference.” To make progress, we can instead estimate the covariate-adjusted difference in average outcomes between treated and control groups:

$$(\text{Covariate-adjusted difference}) E[Y_i | W_i = 1] - E[E[Y_i | W_i = 0, X_i] | W_i = 1].$$

We refer to this as a *statistical estimand*, as opposed to a *causal estimand*, because it can be estimated from observed data when overlap holds, an assumption to which we return below. In LaLonde's context, the statistical estimand represents the difference between the average 1978 earnings of individuals who participated in the program and the average 1978 earnings of nonexperimental individuals with similar observed characteristics.

Unlike a statistical estimand, ATT is a causal estimand because it compares potential outcomes for the same individuals—specifically, those who participated in the program. These two estimands are not necessarily equal. The counterfactual mean for the treated individuals, $E[Y_i(0) | W_i = 1]$, may differ from $E[E[Y_i | W_i = 0, X_i] | W_i = 1]$, a weighted average of observed outcomes for individuals in the nonexperimental control group with similar characteristics. For

³For simplicity, here we do not distinguish between a finite study population and a super-population.

the statistical estimand to recover ATT, a key assumption—unconfoundedness—must hold. That is, if unconfoundedness holds, we can approximate the average untreated outcome for the treated individuals using outcomes from control individuals of similar observed characteristics.

A substantial body of subsequent research has focused on improving methods for estimating the covariate-adjusted difference, particularly in high-dimensional settings. Many recent approaches draw on machine learning techniques to estimate the relationships between covariates, treatment, and outcomes in a flexible manner. Formal results often rely on additional regularity conditions, such as the smoothness of conditional means and propensity scores. For formal treatments, see the references in Imbens and Wooldridge (2009) and Abadie and Cattaneo (2018). The methodological advances in this literature apply well beyond the LaLonde setting.

Our focus here is on the plausibility of the key assumptions. In particular, if unconfoundedness does not hold, we may still be able to estimate the covariate-adjusted difference robustly and precisely, but it cannot be interpreted as a causal effect and may have little substantive relevance.

Unconfoundedness

The unconfoundedness assumption was first explicitly introduced by Rubin (1978) as part of the “ignorable treatment assignment” concept (see also Rosenbaum and Rubin 1983a). If unconfoundedness holds, then treatment assignment can be viewed as effectively random once differences in observed covariates between the treatment and control groups are adjusted for. This assumption plays a central role in identifying causal effects from nonexperimental data. In the context of estimating the average treatment effect on the treated (ATT), unconfoundedness is formally stated as:

$$(\text{Unconfoundedness}) \quad W_i \perp Y_i(0) \mid X_i,$$

which means that treatment status is conditionally independent of the control potential outcome given the covariates. In LaLonde’s context, unconfoundedness means that once we account for observable characteristics (including age, education, race, marital status, and prior earnings), whether someone was in the experimental treatment group or nonexperimental comparison group provides no additional information about what their 1978 earnings would have been had they not participated in the program.

Unconfoundedness is also referred to as conditional independence (Lechner 1999, 2002), or informally as exogeneity (Imbens 2004), or as selection on observables (Barnow, Cain, and Goldberger 1980). However, its definition departs from traditional econometric definitions of exogeneity, which are expressed in terms of residuals and specific functional forms. By contrast, the formal statement of unconfoundedness avoids functional form assumptions, focusing instead on the treatment assignment mechanism. It allows researchers to separate the essence of

the identification assumptions from functional form considerations, emphasizing the role of design rather than the full specification of the data-generating process.

A key result from Rosenbaum and Rubin (1983a) shows that unconfoundedness implies that conditioning on the scalar propensity score is sufficient to remove bias from covariate imbalance. This dimensionality reduction—from the full vector X_i to a single index—has made the propensity score a cornerstone of many modern estimators.

When the parametric model for the conditional expectation of the outcome—such as the earnings equation used in LaLonde (1986)—is correctly specified, unconfoundedness implies a zero conditional mean for the error term. Thus, the results reported in Tables 4 and 5 of the original LaLonde (1986) paper can be viewed as relying on a combination of unconfoundedness and correct functional form assumptions. At the time, however, the nonparametric framing of the identifying assumptions and the emphasis on assignment mechanisms were not yet standard in applied work.

In practice, unconfoundedness is a strong assumption, and its plausibility depends heavily on the context. When the treatment assignment mechanism is poorly understood, this assumption may not be credible. Nonetheless, often researchers can assess its plausibility through supplementary analyses. Tools such as placebo tests and sensitivity analyses can help probe the credibility and robustness of causal claims that rely on unconfoundedness. For a general discussion, see Rosenbaum and Rubin (1983a) and Imbens (2004).

In the next section, we illustrate the use of placebo tests to evaluate unconfoundedness. In the LaLonde setting, the ten covariates are clearly pretreatment variables and should be included in any adjustment strategy. In other settings, however, whether to adjust for a given covariate should be less obvious. Rosenbaum (1984) cautions against adjusting for post-treatment variables, and Cinelli, Forney, and Pearl (2022) offer guidance on selecting from among valid pretreatment covariates for causal inference.

Overlap and Balance

To identify the average treatment effect on the treated (ATT)—and to ensure that the statistical estimand is properly defined—we require an overlap assumption, which states that the propensity score is strictly less than one:

$$(\text{Overlap}) \quad \Pr(W_i = 1 | X_i) < 1.$$

In LaLonde's context, this assumption means that for any individual i assigned to the treatment group with a covariate profile $X_i = x_0$, there must also be individuals in the nonexperimental comparison group with the same profile; otherwise, $\Pr(W_i = 1 | X_i = x_0) = 1$, violating the overlap assumption. Overlap ensures that the weighted average $E[E[Y_i | W_i = 0, X_i] | W_i = 1]$, the second term in the statistical estimand, is well-defined. When overlap is violated, it can be restored by trimming

the treatment group—that is, by removing treated units whose covariate profiles are not represented in the control group.⁴

Overlap implies that treated and control units must share common support in their covariate distributions. Without overlap—for instance, when some covariate values appear in the treatment group but not in the control group—it becomes difficult to make credible comparisons, because estimates rely on extrapolation. Overlap is conceptually distinct from balance, which refers to the similarity in covariate distributions across groups. In a randomized control trial, overlap holds by design, and balance is achieved in expectation. In that case, balance can be further improved through stratification or post-stratification. In nonexperimental settings, ensuring overlap is a key condition for credible estimation of causal effects.

Overlap is especially important when researchers do not wish to impose strong functional form assumptions on the conditional means of potential outcomes or on the structure of treatment effect heterogeneity. When the number of covariates is small, overlap can be assessed by examining marginal or joint covariate distributions across treatment groups. But this becomes impractical in high-dimensional settings. In such cases, it is more effective to inspect the distribution of estimated propensity scores across treated and control groups. Lack of overlap in covariate distributions implies, and is implied by, a lack of overlap in propensity score distributions.

LaLonde did not explicitly discuss overlap, nor did he assess it beyond reporting covariate means by treatment status. Both the regression and selection models he used rely on correct functional form assumptions, which permit interpolation or extrapolation of treatment effects even when treated and control units differ substantially in their covariates—thus bypassing the need for overlap. Still, LaLonde clearly recognized the potential problem: to improve comparability, he trimmed the comparison groups based on “characteristics [that] are consistent with some of the eligibility criteria used to admit applicants into the NSW program” (LaLonde 1986, p. 611). However, by modern standards, his trimming procedures—such as removing all men working in March 1976 in one subset (CPS-SSA-2) or further excluding unemployed respondents with 1975 incomes above the poverty line in another (CPS-SSA-3)—are ad hoc and certainly do not guarantee overlap on all relevant covariates.

Over the past four decades, researchers have developed more systematic approaches for improving overlap, often using the propensity score. These methods vary depending on whether the goal is simply to ensure overlap or to further improve covariate balance. Improving overlap or balance typically requires dropping some units from the sample. Although this reduces the sample size, by ensuring better balance it may improve precision for estimates of the average treatment effect. In addition, the gain in robustness and reduction in bias often outweigh any loss in

⁴ If the target parameter is the average treatment effect (ATE), a stronger version of the overlap assumption is needed: the propensity score must lie strictly between zero and one for all units.

precision. In practice, any increase in variance, even from substantial trimming, is typically modest.⁵

Several specific trimming strategies have been proposed. For instance, focusing on overlap for estimation of the ATT, Dehejia and Wahba (1999a) dropped control units with estimated propensity scores below the minimum observed in the treated group. Crump et al. (2009) proposed a more aggressive approach, selecting subsamples that minimize the variance of average treatment effect (ATE) estimates and recommending trimming units with propensity scores outside the $[0.1, 0.9]$ interval. Crump et al. (2006) and Li, Morgan, and Zaslavsky (2018) proposed improving balance through propensity score weighting, introducing “overlap weights” proportional to the product of the propensity score and one minus the propensity score. Another effective strategy—especially when targeting the ATT—is to match each treated unit to a control unit with a similar estimated propensity score. This approach not only ensures overlap but also tends to improve balance across the covariate distributions.

In practice, overlap—like unconfoundedness—is essential for credible estimation. In settings with poor overlap, such as the LaLonde nonexperimental samples, trimming the sample to ensure overlap is often more important than the choice of estimation method.

Estimation Given Unconfoundedness and Overlap

All estimators in LaLonde (1986) are linear in the covariates. Since then, a wide range of more flexible methods have been proposed to estimate average causal effects under unconfoundedness and overlap. These approaches can be broadly categorized into three groups: (1) outcome modeling, including linear regressions; (2) methods that directly adjust for covariate imbalance, including those based on propensity scores; and (3) doubly robust methods.

First, outcome modeling remains the most widely used approach among applied researchers. It typically involves regressing the outcome on the treatment indicator and covariates (usually, the level terms)—what LaLonde called the earnings equation. This method assumes linearity in covariates and constant treatment effects. A modest relaxation of this approach involves estimating two separate linear regressions for the treated and control groups, sometimes referred to as the Oaxaca-Blinder estimator (Kline 2011). More flexible alternatives include semiparametric and nonparametric methods to model the conditional means of the potential outcomes (for example, Heckman, Ichimura, and Todd 1997; Athey, Tibshirani, and Wager 2019).

⁵ For example, suppose we have a sample with N_t treated units and N_c control units. Under homoskedasticity and random assignment, the variance of the difference-in-means estimator is $\sigma^2(1/N_t + 1/N_c)$. In the LDW-CPS sample, starting with $N_t = 185$ and $N_c = 15,922$, dropping 15,737 control units—a 99 percent reduction—raises the standard error by only about 30 percent in the “best-case scenario,” which assumes no bias from including the additional controls.

The second group of methods focuses on directly adjusting for covariate imbalance between the treatment and control groups. This includes blocking on covariates (that is, grouping units with similar characteristics and comparing outcomes within groups, and then aggregating), covariate matching (for example, Abadie and Imbens 2006, 2008, 2011, 2016; Diamond and Sekhon 2013; Rubin 2006; Imbens 2015), and weighting methods to achieve covariate balance (for example, Hirano, Imbens, and Ridder 2003; Hainmueller 2012; Zubizarreta et al. 2023; Zubizarreta 2015).

Covariate matching is a nonparametric method that avoids imposing modeling assumptions. However, it suffers from the curse of dimensionality when many covariates are present, making it prone to large biases (Abadie and Imbens 2006). In such settings, adjusting for the estimated propensity score is often more practical. This can be implemented through blocking and matching (for example, Dehejia and Wahba 1999a; Abadie and Imbens 2011) or through “inverse propensity weighting” (IPW). IPW reweights observations by the inverse of their estimated propensity score, creating a pseudo-population in which treatment assignment is uncorrelated to observed covariates. Hirano, Imbens, and Ridder (2003) show that a variant (the Hájek variant) of the IPW estimator can achieve the semiparametric efficiency bound even when the propensity score is estimated nonparametrically. In the LaLonde (1986) setting, the IPW estimator reweights the nonexperimental control group based on estimated propensity scores so that its covariate distribution closely approximates that of the experimental treated group.

Because covariate imbalance is the sole source of bias under unconfoundedness, scholars have developed methods to improve balance either by refining propensity score estimation or by bypassing it entirely. For instance, Imai and Ratkovic (2014) propose estimating a covariate-balancing propensity score using the generalized method of moments, whereas Hainmueller (2012) introduces “entropy balancing” to achieve balance in specified covariate moments directly. Entropy balancing can be viewed as an IPW estimator that implicitly relies on a correctly specified propensity score model, where the link function is logistic and the log-odds are linear in covariates (Zhao and Percival 2017).

However, neither outcome modeling nor matching or weighting methods on their own are currently most favored in the methodological literature. Instead, hybrid approaches combine outcome modeling, such as regression, with techniques that address covariate imbalance, such as propensity score weighting, to leverage the strengths of both. Examples include regression within propensity score blocks (Rosenbaum and Rubin 1983b; Imbens 2015), matching followed by regression adjustment (Rubin 1973; Abadie and Imbens 2011), and methods that integrate weighting with regression (for example, Robins, Rotnitzky, and Zhao 1994; Robins and Rotnitzky 1995). These approaches are motivated by the fact that even if balancing or propensity score methods are consistent and efficient in large samples, combining them with outcome modeling can reduce small-sample bias and improve precision. For instance, in high-dimensional settings, the bias from a matching estimator due to remaining covariate imbalance may dominate

the variance, and regression adjustment that accounts for this imbalance can help reduce the bias (Abadie and Imbens 2011).

Robins and Rotnitzky (2001) introduced the term “double robustness,” a key concept for hybrid methods. They show that if either the propensity score or the outcome model is correctly specified, the augmented inverse propensity weighting (AIPW) estimator—which combines propensity score weighting and regression—is consistent. AIPW can be viewed as an outcome model augmented by a correction term: an IPW estimator applied to the residuals from the outcome model, rather than the raw outcomes. The double robustness property arises because, if the outcome model is correctly specified, the correction term has mean zero even if the propensity score model is misspecified; if the outcome model is misspecified, a correctly specified IPW component can debias the regression model. Beyond robustness, AIPW is also desirable for its efficiency: when both models are correctly specified, it achieves the semiparametric efficiency bound (Bang and Robins 2005).

More recently, machine learning methods have become increasingly popular in applied causal inference (Van der Laan and Rose 2011; Wager and Athey 2017; Chernozhukov et al. 2017; Athey, Imbens, and Wager 2018; Athey, Tibshirani, and Wager 2019); see Athey and Imbens (2017, 2019) for reviews. These methods are particularly useful for estimating “nuisance parameters”—such as propensity scores or conditional outcome means—that are not of direct interest but are essential for identifying causal effects. In the LaLonde setting, the goal is to estimate the average treatment effect on the treated (ATT) on post-training wages, which requires modeling the propensity score and/or the conditional means of potential outcomes. Many modern estimators satisfy the “Neyman orthogonality” condition (Chernozhukov et al. 2017), which reduces the effect of small estimation errors in the nuisance parameters on estimates of the target parameter. In binary treatment settings like LaLonde’s, these estimators resemble the AIPW estimator introduced earlier, but with the nuisance parameters estimated via flexible machine learning methods instead of parametric models. Chernozhukov et al. (2017, 2018) emphasize that this property ensures valid inference even when using machine learning algorithms that converge more slowly than is required for estimators based on only estimating the conditional outcome distributions or the propensity score.

Alternative Estimands and Heterogeneous Treatment Effects

Much of the methodological and applied research has focused on estimating average causal effects, such as the average treatment effect on the treated (ATT). However, other quantities may also be of interest. For example, researchers often seek to understand treatment effect heterogeneity among treated units, which can shed light on mechanisms, improve evaluations of effectiveness, and guide personalized policy design. For example, the MDRC (1980) team reported that the National Supported Work program had a large and positive effect on female participants, a significant impact on ex-addict male participants, a small and highly variable impact on ex-criminal male participants, and almost no effect on youth participants.

Econometrically, exploring treatment effect heterogeneity given observed characteristics involves estimating the “conditional average treatment effects on the treated” (CATT). Machine learning methods have been proposed to estimate CATT either nonparametrically or using low-dimensional representations, such as causal forests, while still permitting valid inference or error bounds (for example, Athey and Imbens 2016; Wager and Athey 2018; Athey, Tibshirani, and Wager 2019).

Another important, though less frequently used, estimand is the quantile treatment effects—defined as the difference between quantiles of the treated and untreated potential outcome distributions, either for the population or for the treated group. Under unconfoundedness and overlap, the full marginal distributions of potential outcomes are identified, enabling identification of quantile treatment effects. Firpo (2007) proposes a semiparametrically efficient inverse propensity weighting estimator for these quantities.

Validation through Placebo Analyses

While researchers can assess overlap using observed data, the unconfoundedness assumption is not directly testable. To evaluate the credibility of treatment effect estimates, the literature has developed two main strategies: placebo analyses and sensitivity analyses. Here, we focus on the former and relegate discussion of the latter to the Supplemental Appendix. Placebo analyses offer an indirect way to probe the plausibility of unconfoundedness. These analyses typically involve estimating a model similar to the main specification, but replacing the outcome variable with a pseudo-outcome—usually a variable known to be unaffected by the treatment. A common approach is to test for a treatment effect on a pretreatment variable, such as a lagged outcome, which should not be influenced by the treatment but may still correlate with unobserved confounders. Another variant tests the effect of a pseudo-treatment—often a proxy for the treatment—on the actual outcome. This strategy is often implemented using multiple control groups. For further discussion, see Rosenbaum (1987), Imbens and Rubin (2015), and Imbens (2015).

Although the term “placebo test” was not yet in use in the mid-1980s, LaLonde conducted such an analysis. He regressed 1975 earnings, a pretreatment variable, on the treatment indicator and covariates (reported in columns 2 and 3 of his Tables 4 and 5). Using nonexperimental data, he found that many of the estimated effects were large, negative, and statistically significant, suggesting a violation of unconfoundedness. One limitation of the LaLonde data is that they contain only a single pretreatment outcome. In contrast, the LaLonde-Dehejia-Wahba subset of the LaLonde data allows for placebo analyses that condition on an earlier pretreatment variable, 1974 earnings, when testing for associations between treatment and 1975 earnings. In general, when multiple pretreatment periods are available, placebo tests can be both statistically more powerful and substantively more credible.

Formally, a placebo test assesses a conditional independence restriction, which implies that the placebo outcomes are unrelated to treatment assignment once we account for the remaining observed covariates. A limitation of LaLonde’s original test is that it checks only one implication of the full conditional independence

assumption: whether the average outcomes, after adjusting for observed covariates, are the same across groups. Imbens (2015) discusses other testable implications of this assumption.

Reanalyzing the LaLonde Data

To demonstrate the methodological advances since LaLonde (1986) in practice, we revisit the LaLonde data, including the LaLonde-Dehejia-Wahba data, the original LaLonde male samples, and the LaLonde-Calónico-Smith female samples. Our primary focus is on the LaLonde-Dehejia-Wahba data, while results for the other two datasets are reported in the Supplemental Appendix. For all three datasets, overlap is a central concern. The average treatment effect on the treated (ATT) and conditional average treatment effects on the treated (CATT) are our primary causal estimands.

Propensity Scores and Overlap

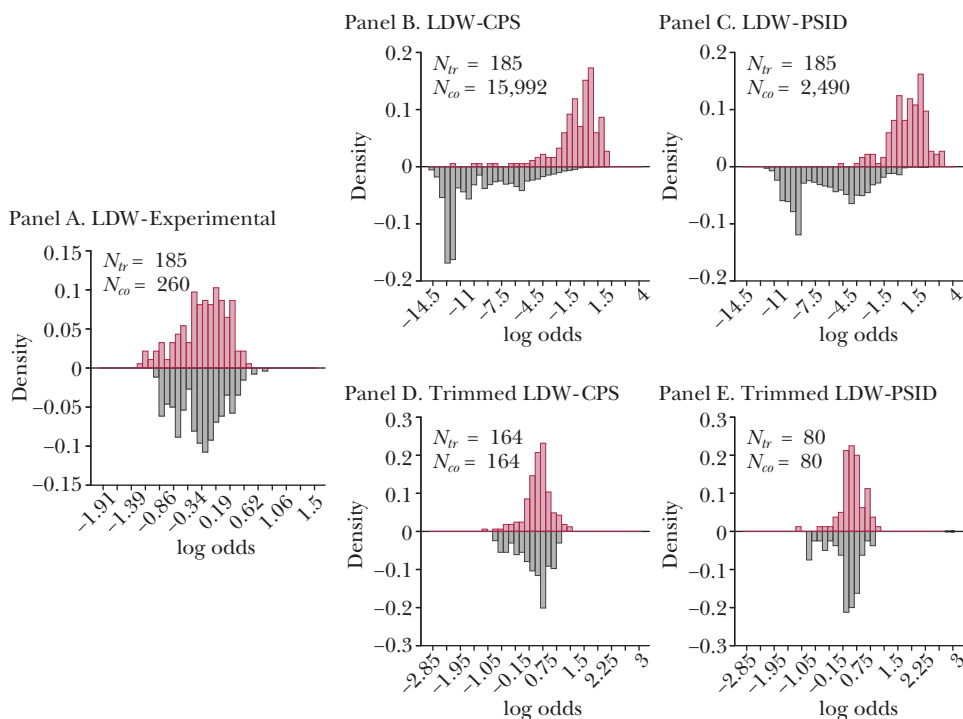
We focus on the LaLonde-Dehejia-Wahba (LDW) data because they include earnings and employment information from 1974. Our analysis uses three LDW datasets: (1) LDW-Experimental, consisting of 185 treated and 280 control individuals from the experimental sample; (2) LDW-CPS, which includes the same treated individuals and 15,922 controls from CPS-SSA-1; and (3) LDW-PSID, comprising the same treated individuals and 2,490 controls from PSID-1. As noted earlier, LaLonde constructed smaller subsamples to improve comparability, but we instead rely on more modern, data-driven methods to assess and address overlap issues.

First, we estimate propensity scores using generalized random forest (Athey et al. 2019), a machine learning method that flexibly models conditional probabilities. Unlike simple logistic regression, generalized random forest can capture complex nonlinearities and higher-order interactions in the covariates, potentially yielding more accurate estimates of the propensity scores.

Panels A–C of Figure 1 use these estimated propensity scores to assess overlap between treated and control units in each sample. Each panel plots histograms of the log-odds of the estimated propensity scores, defined as $\log(\hat{e}/(1 - \hat{e}))$, where \hat{e} is the estimated propensity score. We use the log-odds scale because it more effectively distinguishes differences at the tails of the distribution. For reference, a log-odds of -3 corresponds roughly to a probability of 0.05, and by symmetry, a log-odds of 3 corresponds to a probability of about 0.95.

In panel A, the LDW-Experimental sample shows near-perfect overlap: the treated and control groups have closely aligned propensity score distributions. In contrast, panel B and C reveal severe overlap problems in the nonexperimental samples, with many treated units having propensity scores that fall outside the support of the control group, and large segments of the control group exhibiting extremely low log-odds. Similar patterns are observed in the original LaLonde male samples (as shown in the Supplemental Appendix).

Figure 1

Assessing Overlap in the Lalonde-Dehejia-Wahba (LDW) Data

Source: Imbens and Xu (2025).

Note: Histograms show the log odds ratios for units in the treatment group (top, in maroon) and the control group (bottom, in gray). For unit i , the log odds ratio is defined as $\log(\hat{e}/(1 - \hat{e}))$, where \hat{e} is the estimated propensity score from the generalized random forest. Each figure represents a different sample. N_{tr} and N_{co} represent the numbers of treated and control units, respectively. A: LDW-Experimental. B: LDW-CPS. C: LDW-PSID. D: Trimmed LDW-CPS. E: Trimmed LDW-PSID. For panels D and E, the propensity scores are re-estimated after trimming.

To address these issues, we construct trimmed versions of the LDW-CPS and LDW-PSID samples. Following the earlier discussion, trimming can improve robustness with only modest loss of precision. We begin by augmenting each nonexperimental sample with the experimental controls and estimating each unit's probability of being in the experimental data using the generalized random forest technique. We then trim based on preset thresholds, which may exclude some treated units (Crump et al. 2009). Next, we re-estimate the propensity scores in the trimmed sample and perform 1:1 matching to refine the control group. This yields two sets of trimmed samples: one with experimental treated units and matched nonexperimental controls, and a second with treated and control units from the experiment, which serves as a benchmark. This two-step trimming and matching procedure improves overlap while preserving a comparison to an experimental

benchmark.⁶ As shown in Figure 1, panels D and E, overlap improves substantially in both trimmed samples, albeit at the cost of smaller sample sizes.

Estimating the Average Treatment Effect on the Treated (ATT)

We now estimate the ATT using both the original LaLonde-Dehejia-Wahba nonexperimental samples and the newly constructed trimmed samples. We apply a range of estimators, some likely familiar to most readers and others perhaps more novel. Our primary goal is to compare these methods; again, full computational details are provided in the Supplemental Appendix.

Figure 2 presents the results. The top row reports experimental benchmarks using both the untrimmed and the trimmed LaLonde-Dehejia-Wahba samples. The remaining rows combine the experimental treated units with a nonexperimental control group and apply the following methods: difference-in-means, simple regression; regression with interactions; generalized random forest (GRF) for outcome modeling; “nearest neighbor matching” with bias correction (matching each treated unit with five control units based on covariates); inverse propensity weighting with GRF-estimated propensity scores; covariate balancing propensity score; entropy balancing; double/debiased machine learning using elastic net (DML-ElasticNet); and augmented inverse propensity weighting via GRF (AIPW-GRF). All estimators use the same ten covariates used in prior analyses.⁷

The left panel of Figure 2 shows ATT estimates and 95 percent confidence intervals from the LDW-CPS sample; the right panel shows results from the LDW-PSID sample. For each method, estimates from the full sample are shown in black, and those from the trimmed data are in red. Solid circles mark point estimates; lines represent 95 percent confidence intervals.

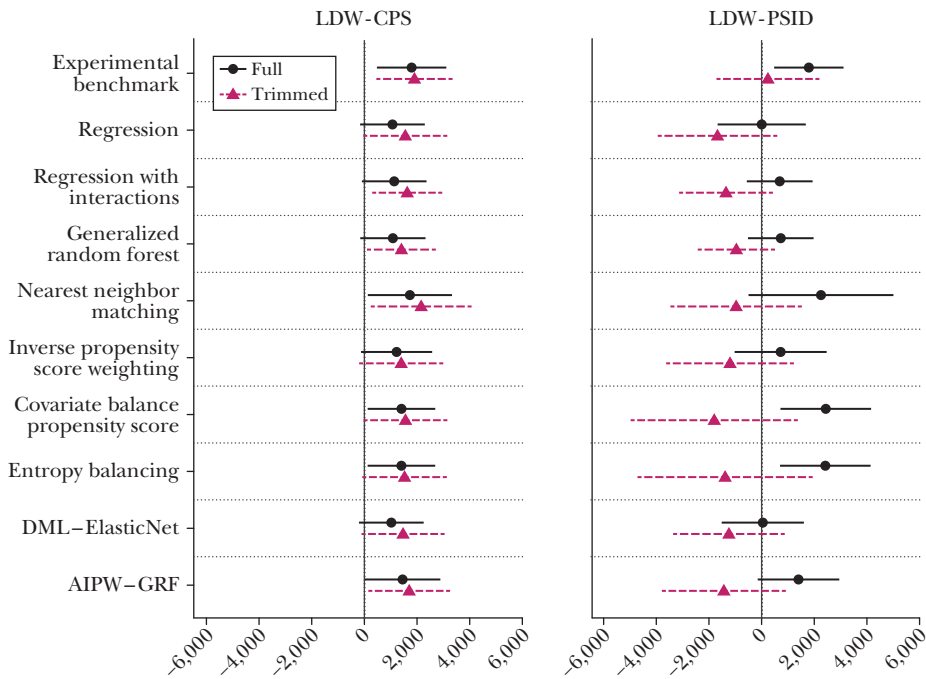
Using the LDW-CPS sample (left panel), all estimators yield positive ATT estimates, though they vary in magnitude. Nearest neighbor matching aligns most closely with the experimental benchmark of \$1,794; covariate balancing propensity score, entropy balancing, and AIPW-GRF also produce estimates near this benchmark. Despite numerical differences, the estimates are not statistically distinguishable from one another. In the trimmed LDW-CPS sample, the estimates show less variation across methods, although confidence intervals widen slightly. All estimates in the trimmed sample center around the experimental benchmark of \$1,911. This stability of estimates after trimming is consistent with the simulation results in Athey et al. (2024).

Using the LDW-PSID sample (right panel), estimates from the full data show more dispersion, ranging from \$4 to \$2,420. AIPW-GRF yields an estimate closest to the experimental benchmark. In the trimmed LDW-PSID sample, the experimental

⁶ Details of the trimming procedure are provided in the Supplemental Appendix along with other computational details.

⁷ The difference-in-means estimates are shown in the Supplemental Appendix but omitted here due to their extreme values in the LDW-CPS and LDW-PSID samples (\$−8,497 and \$−15,204, respectively). In the trimmed samples, they are closer to the other estimates (\$1,483 and \$−1,505).

Figure 2
ATT Estimates Given Unconfoundedness: LaLonde-Dehejia-Wahba (LDW)
Samples



Source: Imbens and Xu (2025).

Note: The panels above show the average treatment effect on the treated (ATT) estimates and their 95 percent confidence intervals using four different samples: LDW-CPS and Trimmed LDW-CPS (left panel), and LDW-PSID and Trimmed LDW-PSID (right panel). Estimates based on corresponding experimental samples are presented at the top. Ten estimators are employed, including difference-in-means, linear regression, linear regression with interactions, generalized random forest for outcome modeling, 1:5 nearest neighbor matching with bias correction, inverse propensity score weighting with GRF-estimated propensity scores, covariate-balance propensity score, entropy balancing, double/debiased machine learning with elastic net (DML-ElasticNet), implemented using DoubleML, and augmented inverse propensity score weighting with GRF for both outcome modeling and propensity score estimation (AIPW-GRF), implemented using generalized random forest techniques.

benchmark is \$306, which is not statistically different from zero at the 5 percent level. Although the estimates from the trimmed sample are all negative and more similar to each other, they seem qualitatively different from the experimental benchmark. However, due to large standard errors, these differences are not statistically significant.

Overall, the results suggest that improving overlap based on observed covariates reduces model dependence and estimate variability, thus yielding more robust estimates of the statistical estimand. However, the fact that many methods produce ATT estimates close to the experimental benchmark using LDW-CPS may have given

researchers a false sense of confidence that modern estimators can recover causal effects—even in settings where the unconfoundedness assumption is likely violated.

Treatment Effect Heterogeneity

We explore treatment effect heterogeneity by estimating the conditional average treatment effects on the treated (CATT), comparing estimates from the experimental and nonexperimental samples in the LaLonde-Dehejia-Wahba data. For simplicity, we focus on the LDW-CPS data—both full and trimmed—for the nonexperimental analyses, using the corresponding experimental samples for benchmarking. In the trimmed sample, 21 treated units lacking comparable controls are removed, and the nonexperimental control group is trimmed using one-to-one matching on the estimated propensity score with the remaining 164 treated units. Details on the trimming procedure are discussed earlier in the paper and in the Supplemental Appendix. We estimate CATT using a causal forest estimator (Athey, Tibshirani, and Wager 2017), with technical details presented in the Supplemental Appendix.

Figure 3 plots the estimated CATT at the covariate values of each treated unit, with experimental data on the x -axis and nonexperimental data on the y -axis, using both the full sample and the trimmed sample. Each dot represents a pair of CATT estimates for an individual with a given covariate profile. The red cross marks the pair of ATT estimates obtained using the AIPW-GRF estimator.

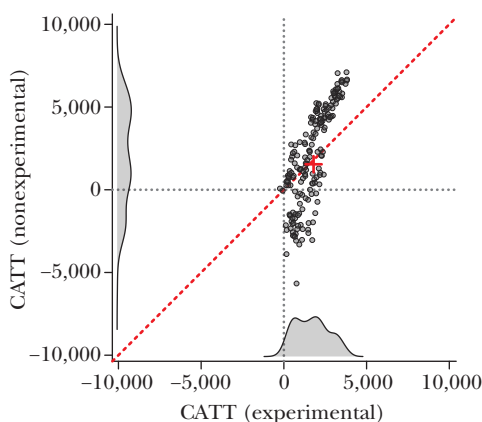
Figure 3 reveals several interesting patterns. First, there is substantial heterogeneity in the estimated treatment effects: CATT estimates based on the experimental data range from \$−236 to \$3,817 in the full sample and from \$−218 to \$4,324 in the trimmed sample. Second, in the full sample, although the AIPW-GRF estimator yields ATT estimates that closely match the experimental benchmark, the CATT estimates from the experimental and nonexperimental data diverge sharply from the 45-degree line. In particular, the nonexperimental CATT estimates span a much wider range (from \$−5,667 to \$7,102) far exceeding that of the experimental estimates, with over one-fourth of treated units obtaining negative CATT estimates. Third, improving overlap substantially enhances the robustness of the CATT estimates: in the trimmed sample, CATT estimates from the experimental and nonexperimental data roughly align along the 45-degree line, and the range of the nonexperimental estimates (from \$−2,877 to \$5,758) is much narrower than in the untrimmed sample.

Applying the same procedure to the LDW-PSID data yields CATT estimates ranging from \$−8,422 to \$4,870 in the full sample and from \$−5,088 to \$1,571 in the trimmed sample—showing greater deviation from the experimental benchmark than in the LDW-CPS case. Analysis of the LaLonde male sample and the reconstructed female data further demonstrates that recovering the CATT is substantially more difficult than estimating the ATT. The combination of the experimental subsample constructed by Dehejia and Wahba (1999a) and the CPS nonexperimental comparison group appears to be an outlier in its ability to recover experimental benchmarks. These additional results are reported in the Supplemental Appendix.

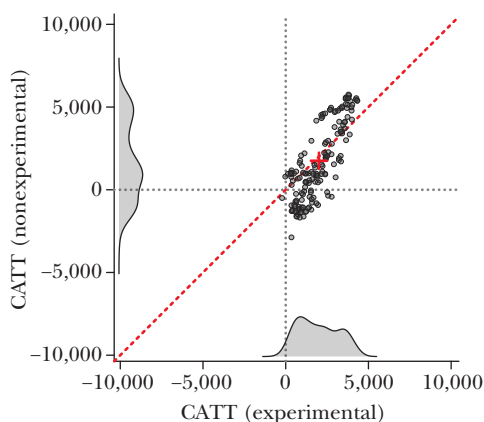
Figure 3

CATT Estimates: Experimental versus Nonexperimental

Panel A. LDW-CPS



Panel B. Trimmed LDW-CPS



Source: Imbens and Xu (2025).

Note: In the above figures, the scatterplots compare conditional average treatment effects on the treated estimates (CATT) from the experimental data (x -axis) and the nonexperimental data (y -axis). Panel A compares CATT estimates from LDW-Experimental versus LDW-CPS, while panel B compares those from their trimmed versions, which exclude individuals with extreme estimated propensity scores. In each figure, a dot represents a CATT estimate based on the covariate values of a treated unit; the red cross marks the pair of ATT estimates from the two samples. The marginal distributions of the CATT estimates are shown as density plots along the axes—the bottom plot for the experimental data and the left-side plot for the nonexperimental data. The red dashed line indicates the 45-degree line. Uncertainty estimates are not shown in this figure.

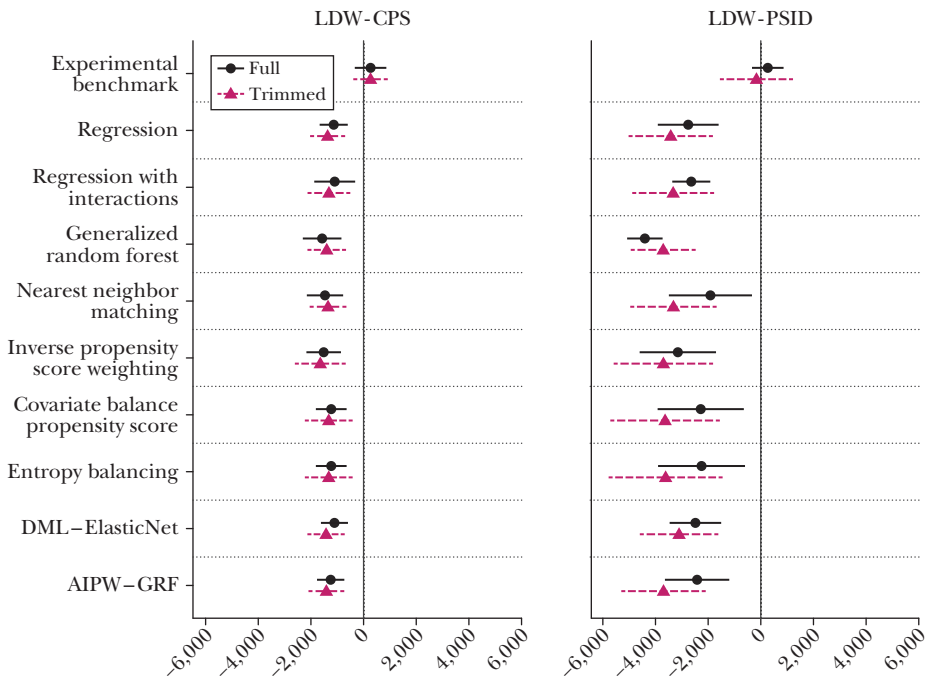
Validation through Placebo Analyses

While modern nonexperimental methods may be effective in estimating the statistical estimand—the covariate-adjusted difference in average outcomes between treated and control groups—this does not imply that the estimate approximates the causal estimand, such as the ATT. Identifying the ATT requires the unconfoundedness assumption, which is fundamentally untestable. However, we can assess its plausibility through placebo analyses.

In the LaLonde setting, we use 1975 earnings as a placebo outcome and exclude both 1975 earnings and employment status from the set of conditioning variables. By construction, 1975 earnings could not have been affected by the treatment, which occurred afterward. We also construct two new trimmed samples, omitting these variables during the trimming process. We then estimate the ATT for the placebo outcome, adjusting for the remaining covariates using a range of estimators.

If the ATT estimates for the placebo outcome are close to zero, this lends support to the unconfoundedness assumption, as it suggests that even without conditioning on 1975 earnings or employment status, treatment assignment is likely independent of the untreated potential outcome $Y_i(0)$ —that is, 1978 earnings had

Figure 4
Placebo Tests: 1975 Earnings as the Outcome



Source: Imbens and Xu (2025).
Note: The figures above show the placebo estimates and their 95 percent confidence intervals using four samples: LDW-CPS and Truncated LDW-CPS (left panel), and LDW-PSID and Truncated LDW-PSID (right panel). For each case, results based on the corresponding experimental samples are shown at the top as benchmarks. All estimates are produced using the same ten estimators introduced earlier.

the individual not participated in the program. Including these variables in the main analysis would then make unconfoundedness even more credible. However, if the placebo ATT estimates differ significantly from zero, this may indicate either that 1975 earnings and employment status are key confounders, or that unmeasured factors such as perseverance influence both treatment assignment and outcomes. In that case, the placebo analysis fails to bolster unconfoundedness in the main analysis.

Figure 4 presents the results. As expected, the experimental benchmarks are close to zero and statistically insignificant. In contrast, all estimators using nonexperimental data yield large, negative estimates. While trimming improves the stability of these estimates, they remain statistically different from zero.⁸ Moreover, as shown

⁸ The difference-in-means estimates are not shown here; they are \$-12,118, \$-17,531, \$-14,056, and \$-4,670 across the four panels, respectively. Details are available in the Supplemental Appendix.

in the Supplemental Appendix, CATT estimates for a similar placebo test based on experimental data cluster around zero, while their nonexperimental counterparts are consistently negative and sizable, indicating substantial bias. These patterns further illustrate the point we aim to emphasize: With sufficient overlap, modern methods can robustly estimate the statistical estimands, but not necessarily the causal estimands if unconfoundedness is violated.

Alternative Samples

For comparison, we also revisit the original male samples used in LaLonde (1986). These datasets lack information on 1974 earnings and employment status. We find that with sufficient overlap, most modern estimators yield estimates within relatively narrow ranges when using CPS-SSA-1 or PSID-1 as control groups. However, these estimates—most of which are negative—do not align with the experimental ATT benchmarks. Similar patterns are reported by Smith and Todd (2001, 2005). Using these nonexperimental data, modern methods also fail to recover the experimental CATT benchmarks, as noted earlier.

Using the LaLonde-Calónico-Smith female sample with PSID data as nonexperimental controls (Calónico and Smith 2017), we find that many modern methods produce estimates close to the experimental benchmarks, though standard errors are often quite large. As noted by Calónico and Smith (2017), selection appears to be less severe for female participants than for male participants in the National Supported Work program. However, overlap remains a significant challenge. Moreover, a placebo test using the number of children in 1975—a variable not included in LaLonde's original analysis—does not support the unconfoundedness assumption. CATT estimates also do not align with the experimental benchmarks.

As an example of a dataset that readily passes a placebo test, we include in the Supplemental Appendix a reanalysis of data from Imbens, Rubin, and Sacerdote (2001), who conducted an original survey to study the impact of lottery prize size in Massachusetts during the mid-1980s on the economic behavior of lottery players. The primary outcome is post-winning labor earnings, with data available for earnings in the six preceding years. While lottery outcomes may seem random, there are systematic differences in pretreatment variables, likely related to the number of tickets purchased or survey response rates. However, in this case, placebo tests provide strong evidence supporting the unconfoundedness assumption, bolstering the credibility of the causal estimates. The availability of six pretreatment earnings measures proves particularly valuable: they likely capture both selection and outcome-relevant factors and serve as strong placebo outcomes due to their comparability to the post-treatment outcome. We provide the data (Imbens, Rubin, and Sacerdote 2001) in the replication materials.

Summary

Our reexamination of the LaLonde data shows that when overlap is ensured, the choice of estimation method becomes less consequential, as most methods yield similar results. However, these estimates may not represent the causal

quantity of interest if the unconfoundedness assumption is violated. Supplementary analyses, such as placebo tests, can help assess how plausible this assumption is. Although in both the LDW-CPS and LaLonde-Calónico-Smith datasets, many modern methods appear to recover the experimental benchmark for the ATT, placebo tests fail to support unconfoundedness. In typical research settings where experimental benchmarks are unavailable, we cannot know whether such estimates credibly recover the causal estimand. Indeed, with the original LaLonde male sample or the LDW-PSID dataset, modern estimators fail to recover the ATT.

The answer to whether modern nonexperimental methods can credibly estimate causal effects, 40 years after LaLonde's critique, is therefore nuanced. Researchers can now use data-driven approaches to improve overlap and apply modern methods to reliably estimate the statistical estimand. However, credible causal inference still depends on the validity of the unconfoundedness assumption. If supplementary analyses—such as placebo tests—support this assumption, modern methods can yield highly credible estimates. Otherwise, they cannot.

Lessons Learned

What lessons has the methodological literature since LaLonde (1986) taught us? And what specific analyses should researchers conduct today in similar nonexperimental studies? We return to the five lessons mentioned in the introduction and offer practical recommendations.

First, any analysis of causal effects using nonexperimental data should begin with a careful investigation of the treatment assignment mechanism. A clear understanding of the “design” is essential for evaluating the plausibility of the unconfoundedness assumption. The case for relying on unconfoundedness is strongest when researchers believe that selection into treatment is driven by factors that are well understood, observed, and measured. When that condition holds, flexibly adjusting for observed pretreatment covariates can reduce reliance on strong modeling assumptions. In cases where important confounders are unobserved, researchers may turn to panel data methods that account for time-invariant unobserved confounding (for recent overviews, see Xu 2023; Arkhangelsky and Imbens 2024).

Second, the literature has underscored the importance of assessing and improving overlap in covariate distributions. The comparison groups LaLonde used differed substantially from the experimental sample, prompting him to discard some units based on age, employment status, and earnings. Since then, more systematic and data-driven approaches have been developed to diagnose and address lack of overlap, including propensity score-based trimming and weighting strategies. The loss of efficiency is often modest and a worthwhile cost.

Third, relatedly, the propensity score has become central both to diagnosing overlap and to estimating treatment effects. Researchers now routinely estimate the propensity score using flexible methods and evaluate overlap by comparing the

distribution of propensity scores across treated and control groups. When necessary, samples can be trimmed to improve comparability. While the concept of the propensity score was introduced just prior to LaLonde's work in Rosenbaum and Rubin (1983a), it has since become foundational. Doubly robust estimators, particularly those that combine outcome modeling with inverse propensity score weighting, yield consistent estimates when either the outcome model or the propensity score model is correctly specified. The integration of machine learning techniques into causal inference has further strengthened these methods by reducing dependence on ad hoc specification choices. We expect these methods to see broader adoption among economists and social scientists.

Fourth, attention has shifted from solely estimating average treatment effects to understanding treatment effect heterogeneity. Policymakers increasingly ask not only "Does it work?" but "For whom does it work?" and "Where does it cause harm?" The literature has developed tools to estimate conditional average treatment effects and quantile treatment effects. These estimands help unpack heterogeneous impacts and can inform personalized policy decisions. Large datasets and algorithmic advances have made it easier to estimate these effects flexibly and at scale (for example, Wager and Athey 2018). We encourage the estimation and visualization of these additional quantities of interest.

Finally, the credibility of causal estimates increasingly hinges on validation exercises—especially placebo tests. While LaLonde included some placebo analyses using lagged earnings, his main focus was on comparing nonexperimental estimates to experimental benchmarks. In contrast, modern practice places greater emphasis on formal diagnostic checks. Placebo tests, such as estimating effects on outcomes known to be unaffected by the treatment, offer informative checks on key identifying assumptions like unconfoundedness and should be more routinely used in empirical analyses.

To help researchers apply these lessons in practice, we provide a detailed online tutorial—including R code and the data used in this paper and the Supplemental Appendix—available at <https://yiqingxu.org/tutorials/lalonde/>. The tutorial replicates our analyses and can be easily adapted to other datasets.

■ *We thank the Office of Naval Research for support under grant numbers N00014-17-1-2131 and N00014-19-1-2468, and Amazon for a gift. We are grateful to Susan Athey, Scott Cunningham, Alexis Diamond, Peng Ding, Dean Eckles, and Xiang Zhou for helpful comments and feedback, and to Timothy Taylor, Jonathan Parker, and Heidi Williams for detailed editorial suggestions. We also thank Zihan Xie and Jinwen Wu for excellent research assistance.*

References

- Abadie, Alberto, and Matias D. Cattaneo. 2018. "Econometric Methods for Program Evaluation." *Annual Review of Economics* 10: 465–503.
- Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74 (1): 235–67.
- Abadie, Alberto, and Guido W. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76 (6): 1537–57.
- Abadie, Alberto, and Guido W. Imbens. 2011. "Bias-Corrected Matching Estimators for Average Treatment Effects." *Journal of Business and Economic Statistics* 29 (1): 1–11.
- Abadie, Alberto, and Guido W. Imbens. 2016. "Matching on the Estimated Propensity Score." *Econometrica* 84 (2): 781–807.
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.
- Arkhangelsky, Dmitry, and Guido Imbens. 2024. "Causal Models for Longitudinal and Panel Data: A Survey." *Econometrics Journal* 27 (3): C1–61.
- Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 60 (1): 47–57.
- Ashenfelter, Orley, and David Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics* 67 (4): 648–60.
- Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–60.
- Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32.
- Athey, Susan, and Guido W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11: 685–725.
- Athey, Susan, Guido W. Imbens, Jonas Metzger, and Evan Munro. 2024. "Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations." *Journal of Econometrics* 240 (2): 105076.
- Athey, Susan, Guido W. Imbens, and Stefan Wager. 2018. "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80 (4): 597–623.
- Athey, Susan, Julie Tibshirani, Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47 (2): 1148–78.
- Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61 (4): 962–73.
- Barnow, Burt S., Glen G. Cain, and Arthur S. Goldberger. 1980. "Issues in the Analysis of Selectivity Bias." University of Wisconsin-Madison Institute for Research on Poverty Discussion Paper 600-80.
- Calónico, Sebastian, and Jeffrey Smith. 2017. "The Women of the National Supported Work Demonstration." *Journal of Labor Economics* 35 (S1): S65–97.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review* 107 (5): 261–65.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21 (1): C1–68.
- Chernozhukov, Victor, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. 2024. "Applied Causal Inference Powered by ML and AI." Preprint, arXiv. <https://doi.org/10.48550/arXiv.2403.02467>.
- Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2022. "A Crash Course in Good and Bad Controls." *Sociological Methods and Research* 53 (3): 1071–1104.
- Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2006. "Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the

- Estimand.” NBER Working Paper 0330.
- Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik.** 2009. “Dealing with Limited Overlap in Estimation of Average Treatment Effects.” *Biometrika* 96 (1): 187–99.
- Cunningham, Scott.** 2021. *Causal Inference: The Mixtape*. Yale University Press.
- Currie, Janet, Henrik Kleven, and Esmée Zwiers.** 2020. “Technology and Big Data Are Changing Economics: Mining Text to Track Methods.” *AEA Papers and Proceedings* 110: 42–48.
- Dehejia, Rajeev H., and Sadek Wahba.** 1999a. “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association* 94 (448): 1053–62.
- Dehejia, Rajeev H., and Sadek Wahba.** 1999b. *Reconstructed National Supported Work Demonstration Data*. <https://users.nber.org/~rdehejia/data/.nswdata2.html>.
- Dehejia, Rajeev H., and Sadek Wahba.** 2002. “Propensity Score-Matching Methods for Nonexperimental Causal Studies.” *Review of Economics and Statistics* 84 (1): 151–61.
- Diamond, Alexis, and Jasjeet S. Sekhon.** 2013. “Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies.” *Review of Economics and Statistics* 95 (3): 932–45.
- Ding, Peng.** 2024. *A First Course in Causal Inference*. CRC Press.
- Firpo, Sergio.** 2007. “Efficient Semiparametric Estimation of Quantile Treatment Effects.” *Econometrica* 75 (1): 259–76.
- Fraker, Thomas, and Rebecca Maynard.** 1987. “The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs.” *Journal of Human Resources* 22 (2): 194–227.
- Hainmueller, Jens.** 2012. “Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies.” *Political Analysis* 20 (1): 25–46.
- Heckman, James J.** 1978. “Dummy Endogenous Variables in a Simultaneous Equation System.” *Econometrica* 46 (4): 931–59.
- Heckman, James J., and V. Joseph Hotz.** 1989. “Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training.” *Journal of the American Statistical Association* 84 (408): 862–74.
- Heckman, James J., V. Joseph Hotz, and Marcelo Dabos.** 1987. “Do We Need Experimental Data to Evaluate the Impact of Manpower Training on Earnings?” *Evaluation Review* 11 (4): 395–427.
- Heckman, James J., Hidehiko Ichimura, and Petra E. Todd.** 1997. “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme.” *Review of Economic Studies* 64 (4): 605–54.
- Hirano, Keisuke, Guido W. Imbens, and Geert Ridder.** 2003. “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score.” *Econometrica* 71 (4): 1161–89.
- Holland, Paul W.** 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81 (396): 945–60.
- Huber, Martin.** 2023. *Causal Analysis: Impact Evaluation and Causal Machine Learning with Applications in R*. MIT Press.
- Huntington-Klein, Nick.** 2021. *The Effect: An Introduction to Research Design and Causality*. CRC Press.
- Imai, Kosuke, and Marc Ratkovic.** 2014. “Covariate Balancing Propensity Score.” *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76 (1): 243–63.
- Imbens, Guido W.** 2004. “Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review.” *Review of Economics and Statistics* 86 (1): 4–29.
- Imbens, Guido W.** 2015. “Matching Methods in Practice: Three Examples.” *Journal of Human Resources* 50 (2): 373–419.
- Imbens, Guido W., and Donald B. Rubin.** 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Imbens, Guido W., Donald B. Rubin, and Bruce I. Sacerdote.** 2001. “Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players.” *American Economic Review* 91 (4): 778–94.
- Imbens, Guido W., and Jeffrey M. Wooldridge.** 2009. “Recent Developments in the Econometrics of Program Evaluation.” *Journal of Economic Literature* 47 (1): 5–86.
- Imbens, Guido W., and Yiqing Xu.** 2025. *Data and Code for: "Comparing Experimental and Nonexperimental Methods: What Lessons Have We Learned Four Decades after LaLonde (1986)?"* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. <https://doi.org/10.3886/E232201V1>.

- Kline, Patrick. 2011. "Oaxaca-Blinder as a Reweighting Estimator." *American Economic Review* 101 (3): 532–37.
- LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (4): 604–20.
- Lechner, Michael. 1999. "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification." *Journal of Business and Economic Statistics* 17 (1): 74–90.
- Lechner, Michael. 2002. "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies." *Review of Economics and Statistics* 84 (2): 205–20.
- Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. 2018. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association* 113 (521): 390–400.
- MDRC (Manpower Development Research Corporation). 1980. *Summary and Findings of the National Subsidized Work Demonstration*. Ballinger Publishing Company.
- Robins, James M., and Andrea Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90 (429): 122–29.
- Robins, James M., and Andrea Rotnitzky. 2001. Comments on "Inference for Semiparametric Models: Some Questions and an Answer." *Statistica Sinica* 11 (4): 920–36.
- Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66.
- Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society Series A: Statistics in Society* 147 (5): 656–66.
- Rosenbaum, Paul R. 1987. "The Role of a Second Control Group in an Observational Study." *Statistical Science* 2 (3): 292–306.
- Rosenbaum, Paul R., and Donald B. Rubin. 1983a. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.
- Rosenbaum, Paul R., and Donald B. Rubin. 1983b. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society Series B: Methodological* 45 (2): 212–18.
- Rubin, Donald B. 1973. "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics* 29 (1): 185–203.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.
- Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6 (1): 34–58.
- Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. Cambridge University Press.
- Smith, Jeffrey A., and Petra E. Todd. 2001. "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods." *American Economic Review* 91 (2): 112–18.
- Smith, Jeffrey A., and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–53.
- Splawa-Neyman, Jerzy. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay On Principles. Section 9." *Statistical Science* 5 (4): 465–72.
- Van der Laan, Mark J., and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer.
- Wager, Stefan. 2024. "Causal Inference: A Statistical Learning Approach." Unpublished.
- Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42.
- Xu, Yiqing. 2023. "Causal Inference with Time-Series Cross-Sectional Data: A Reflection." In *Oxford Handbook of Engaged Methodological Pluralism*, edited by Janet M. Box-Steffensmeier, Dino P. Christenson, and Valeria Sinclair-Chapman. Oxford University Press.
- Zhao, Qingyuan, and Daniel Percival. 2017. "Entropy Balancing Is Doubly Robust." *Journal of Causal Inference* 5 (1): 20160010.
- Zubizarreta, José R. 2015. "Stable Weights That Balance Covariates for Estimation with Incomplete Outcome Data." *Journal of the American Statistical Association* 110 (511): 910–22.
- Zubizarreta, José R., Elizabeth A. Stuart, Dylan S. Small, and Paul R. Rosenbaum. 2023. *Handbook of Matching and Weighting Adjustments for Causal Inference*. CRC Press.

