LaLonde (1986) after Nearly Four Decades: Lessons Learned*

Guido W. Imbens and Yiqing Xu†

In 1986, Robert LaLonde published a paper based on part of his PhD thesis (LaLonde 1986), which has profoundly impacted both methodological and empirical literatures on estimating causal effects. As of August 2024, this paper has been cited roughly 3,000 times, a number that only partially reflects its tremendous impact on causal inference field and the credibility revolution (Angrist and Pischke 2010). In his paper, he assessed whether the then state-of-the-art nonexperimental evaluation methods could match experimental benchmarks. LaLonde's conclusion was ultimately negative regarding the credibility of the range of nonexperimental methods he examined. He wrote:

> "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." (LaLonde 1986, Abstract, p. 604),

and concluded that

*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.

†*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, tutorials, and author disclosure statements, see the article page at https://doi.org/XXXX

> "policymakers should be aware that the available nonexperimental evaluations of employment and training programs may contain large and unknown biases resulting from specification errors." (LaLonde 1986, Conclusion, p. 617),

A year earlier, LaLonde's thesis advisors Orley Ashenfelter and David Card had raised similar concerns:

> "we conclude that randomized clinical trials are necessary to reliably determine program effects." (Ashenfelter and Card 1985, Abstract, p. 648)

Quickly after the publication of LaLonde's paper, several other studies echoed his concern about the credibility of nonexperimental methods in the context of the evaluation of labor market programs (Fraker and Maynard 1987; Heckman et al. 1987), leading to further skepticism about their value. A couple of years earlier Leamer (1983) had raised concerns about the credibility of applied econometric methods more generally, and had asked "Is Randomization Essential?" (Leamer 1983, p. 31).

In this article, we summarize our perspective on the lessons from the methodological literature that followed LaLonde's seminal paper. We argue that while some of LaLonde's original conclusions have stood the test of time, substantial progress has been made in the methodological literature both in terms of improved estimators and in the form of suggested additional analyses to assess the credibility of the primary analyses. At this point, almost four decades after the publication of LaLonde's paper, the answer to his original question—whether nonexperimental methods can successfully replicate experimental benchmarks—is more nuanced than his original conclusions. First, it is clear now that sometimes we can, and we currently have better methods both for achieving this when we can (with the methods used in the original paper now largely discarded). Second, we now have methods for telling whether we can.

One important development that greatly increased the influence of the original LaLonde study was that Rajeev Dehejia and Sadek Wahba (Dehejia and Wahba 1999, 2002) made the male subsample of the original data widely available.[2] More recently, Calónico and Smith (2017) have also reconstructed the female samples used in LaLonde study. These public datasets are highly valuable for teaching and future research.[3] Throughout the remainder of the paper, we refer to the data compiled by Dehejia and

[2]As part of a research project starting in a graduate class taught by Imbens and Rubin at Harvard in 1996, Dehejia and Wahba obtained the original data from LaLonde, stored on tapes. They successfully located an old tape reader capable of retrieving the data. The data are now publicly available on Dehejia's website (https://users.nber.org/~rdehejia/data/.nswdata2.html) and have been extensively used in the causal inference literature on causal inference.

[3]Using these public datasets, we create a detailed online tutorial to assist readers in implementing the procedures discussed in this article. See https://yiqingxu.org/tutorials/lalonde/.

Wahba as the LaLonde-Dehejia-Wahba data and the reconstructed female samples as the Lalonde-Calónico-Smith data.

Here are the five lessons we see as emerging from this literature. The first lesson underscores the key role of the unconfoundedness assumption, which underpins most modern covariate adjustment methods. This assumption implies that treatment assignment is as if randomly assigned conditional on the observed covariates. The second lesson emphasizes the importance of improving overlap—which captures how similar the covariate distributions of the treatment and control groups are—before estimating treatment effects. The third lesson highlights the central role of the propensity score, defined as the probability of being treated given observed covariates.[4] While the propensity score had barely reached the economics literature when LaLonde (1986) was writing his thesis, it has since become a valuable tool not only for assessing overlap but also as a crucial component of many modern estimators based on unconfoundedness. The fourth lesson is the need to go beyond average causal effects and further investigate effect heterogeneity, as this helps researchers gain a deeper understanding of how the treatment works. Finally, validation exercises, such as placebo tests, for assessing the critical assumptions are crucial for establishing the credibility of research. [5]

To illustrate these lessons in practice, we reexamine the LaLonde data, including the Lalonde-Dehejia-Wahba data, the original LaLonde male samples, and the Lalonde-Calónico-Smith data. We also analyze a second data set where these methods are natural, the Imbens-Rubin-Sacerdote lottery data from Imbens et al. (2001). We show that, once sufficient overlap is ensured, various methods can produce similar and robust estimates for the statistical estimand, that is, the covariate-adjusted difference in the average outcomes between the treatment and control groups. However, these estimands lack a causal interpretation if the unconfoundedness assumption is violated. To assess its credibility, validation exercises such as placebo tests are critical, whereas goodness-of-fit tests are largely irrelevant. For the LaLonde data sets, placebo estimates do *not* support the unconfoundedness assumption. However, optimism is warranted: analyzing the Imbens-Rubin-Sacerdote data, which benefit from a clearer assignment mechanism (lottery) and more detailed pretreatment information, shows that placebo analyses do

[4]A key paper in this literature, Rosenbaum and Rubin (1983b), published just prior to LaLonde's paper, has the title "The Central Role of the Propensity Score in Observational Studies for Causal Effects."

[5]We focus on five issues related to the particular setting studied in LaLonde (1986), the evaluation of an intervention at the individual level, based on detailed background information on those individuals. There is a literature on causal inference more broadly which has grown substantially over the nearly four decades as well, as documented in Currie et al. (2020). We do not cover this broader literature here. There are more comprehensive surveys of the general causal literature, (e.g., Abadie and Cattaneo 2018; Imbens and Wooldridge 2009), as well as numerous textbooks (Angrist and Pischke 2008; Imbens and Rubin 2015; Cunningham 2018; Huntington-Klein 2021; Huber 2023; Ding 2024; Wager 2024; Chernozhukov et al. 2024). Here, we limit the discussion to five lessons learned from the subsequent literature for settings similar to those in LaLonde (1986).

support the unconfoundedness assumption.

LaLonde's Findings

We first describe the data Lalonde used and explain the main econometric approaches in LaLonde (1986). We then examine the regression and selection model estimates in his paper. After summarizing Lalonde's findings, we proceed to describe the Lalonde-Dehejia-Wahba data and explore some of the immediate responses to LaLonde's findings.

LaLonde's Data

LaLonde (1986) analyzed the training effects of the National Supported Work Demonstration (NSW) program for female and male participants separately. The female participants were drawn from the Aid to Families with Dependent Children (AFDC) program. The male participants came from three other target groups, all with extremely poor labor market prospects: ex-drug addicts, ex-criminal offenders, and high-school dropouts. For both female and male participants, LaLonde used two main data sources to construct comparison groups: CPS-SSA-1, drawn from Westat's Matched Current Population Survey–Social Security Administration File, includes all females or males under 55 meeting Westat's criteria; PSID-1, from the Panel Study of Income Dynamics, includes all female or male household heads under 55 from 1975 to 1978 for males and 1979 for females who did not identify as retired in 1975. The age cutoff was chosen to make the comparison group more comparable to the experimental sample.

LaLonde further refined these datasets based on criteria like employment status, time of survey, and poverty status, creating four additional comparison groups: CPS-SSA-2, CPS-SSA-3, PSID-2, and PSID-3. Our subsequent reanalysis will primarily focus on the Lalonde-Dehejia-Wahba male samples because they are widely used in methodological literature and because they include two pretreatment outcomes, earnings in 1974 and earnings in 1975. Results based on the female samples are provided in the online appendix.

Table 1 columns 1–4 provide summary statistics for the male samples in the experimental data and comparison groups used in LaLonde (1986). LaLonde showed that respondents in the 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 of the NSW experiment; they also had substantially higher earnings in years before the program took place. The large difference in average earnings between the comparison groups and the experimental sample (on average, 1975 earnings were $3K in the experimental data, versus $14K in the

TABLE 1.
Descriptive Statistics: LaLonde and Lalonde-Dehejia-Wahba Male Samples

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

Note: Standard deviations are in the parentheses. Tables 3, 5, and 6 in LaLonde (1986) use data described in columns 1–4. Dehejia and Wahba (1999) primarily use data described in columns 3–6.

CPS-SSA-1 sample and $19K in the PSID-1 sample) motivated LaLonde to construct the other comparison groups, using earnings-based criteria. Note that the male CPS-SSA-1 sample, with 15,922 observations, is significantly larger than the male PSID-1 sample, which has 2,490 observations.

The experimental treated units in the Lalonde-Dehejia-Wahba data are a subset of the treated units in the original LaLonde data, specifically chosen to include information on 1974 earnings. Using this subset turns out to be important as it allows analysts to adjust for longer earnings histories and conduct additional placebo analyses.

Econometric Approaches

To estimate the causal effect of the NSW program on 1978 earnings using experimental and nonexperimental data, LaLonde (1986) employed a variety of models, which can be broadly divided into two categories: regression methods that rely solely on an outcome model, referred to as the "earnings equation," and selection models that include an additional "participation equation."

Regression estimates. In Tables 4 and 5 (attached in the appendix), LaLonde (1986) presented results for the training effects on female and male participants, respectively,

using seven different estimators based on six comparison groups, with minor specification variations in the regression function for earnings. All regression models are linear and assume that the error term has zero conditional mean, implying that the regressors are uncorrelated with the error term and, therefore, exogenous. They also implicitly assume a constant treatment effect. The seven models include (i) a simple regression estimator both without or with controls for age, education, and race (but, notably, *not* including earnings in 1975 in the set of controls); (ii) a difference-in-differences estimator using 1975 earnings as the pretreatment outcome—hence, the original outcome is replaced with a first-differenced outcome—both without or with controls for age; (iii) a quasi-difference-in-differences estimator that places 1975 earnings on the right-hand side of the regression to account for transitory shocks, also known as the Ashenfelter dip (Ashenfelter 1978), both without or with controls for age, education, and race; and (iv) a specification that controls for all pretreatment covariates, including pre-training earnings and unemployment status in 1975, as well as marital status. The comparison groups, as described above, include CPS-SSA-1 and PSID-1, as well as subsets thereof with different selection criteria to make them more similar to the NSW sample. Interestingly, and somewhat anticipating the emphasis these analyses would receive in contemporary literature, LaLonde also reported results from placebo tests using 1975 earnings as the placebo outcome.

LaLonde's findings based on the linear regression model estimates are several-fold. First, using the experimental NSW data, all seven estimators produce similar training effect estimates around $851 for female participants and $886 for male participants, and the estimated placebo effects are close to zero. Second, when using nonexperimental comparison groups, the estimates diverge significantly from the experimental benchmarks, often yielding large and negative values in both female and male samples with small to modest standard errors. Third, these estimates vary widely, and specification tests focusing on goodness-of-fit are unlikely to guide an analyst to the experimental benchmarks. Collectively, these findings led LaLonde to conclude that none of the regression adjustment methods popular at the time of his writing, when applied to nonexperimental data, were credible.

Selection model estimates. In addition to these estimates based on exogeneity of the treatment, LaLonde (1986) also presented results in Table 6 (attached in the appendix) that account for endogeneity of the treatment indicator. This approach uses the two-step estimator proposed by Heckman (1978), which allows the error terms in the earnings equation and participation equation to be correlated. Identification relies on the presence of covariates included (with non-zero coefficients) in the participation equation but excluded from the earnings equation, or on an assumption of a joint normal distribution for the error terms. For both female and male samples, LaLonde employed

three comparison groups: the experimental controls, CPS-SSA-1, and PSID-1, and tested four different specifications for females and three for males. Each specification uses a distinct set of variables in the participation equation that are excluded from the earnings equation. *A priori*, none of the specifications appears more justified by economic or econometric theories than any other.

LaLonde reported that all selection model estimates using the experimental data remain close to $851 for females and $886 for males. However, those based on nonexperimental data vary substantially across specifications and, as in the exogenous regression case, deviate substantially from the experimental benchmarks. He concluded that while the two-step procedure brings estimates closer to the experimental benchmarks, it still results in a "considerable range of imprecise estimates" (LaLonde 1986, p. 617).

The LaLonde-Dehejia-Wahba Data

Dehejia and Wahba (1999) focused solely on male participants, stating that "estimates for this group were the most sensitive to functional-form specification" (Dehejia and Wahba 1999, p. 1054). They constructed a subsample from LaLonde's original data that includes participants with available 1974 earnings and unemployment status. Dehejia and Wahba (1999) argued that this subsample remains a valid experimental sample because its construction relies on pretreatment information only, such as month of assignment and employment history, ensuring that treatment assignment remains orthogonal to all pretreatment variables. Notably, this subsample contains only 62% of the original treated group used by LaLonde. They also use the subsets of the same six datasets as LaLonde for nonexperimental controls, which likewise contain 1974 earnings and unemployment information. This collection of datasets, referred to as the LaLonde-Dehejia-Wahba data, is now widely used in causal inference literature.[6]

Columns 5 and 6 in Table 1 show summary statistics for the treated observations in the Lalonde-Dehejia-Wahba data. The table shows that the NSW participants in the Lalonde-Dehejia-Wahba treated sample had a higher unemployment rate in 1975 and lower average earnings in 1975 compared to LaLonde's original male samples. The 1974 earnings data, available only in the Lalonde-Dehejia-Wahba sample, suggest that many Lalonde-Dehejia-Wahba participants faced long-term unemployment. These factors may explain why the estimated training effect in the Lalonde-Dehejia-Wahba sample, $1,794, is more than double that in LaLonde's male sample, which is $886.

[6]In fact, most methodological research on causal inference uses a specific sample of the Lalonde-Dehejia-Wahba data, with controls from CPA-SSA-1.

Subsequent Literature

The publication of LaLonde (1986) ignited a vigorous debate in the applied econometrics literature. Heckman and Hotz (1989) responded to Lalonde's criticism of the nonexperimental evaluation literature by suggesting the use of specification tests to eliminate particularly poor estimators. However, this approach falls short in differentiating among numerous estimators that adequately fit the data but are based on varying identifying assumptions. As a result, this approach has not found many followers in the subsequent literature.

Dehejia and Wahba (1999) introduced propensity score-based stratification and matching methods to address LaLonde's challenge. They obtained estimates close to the experimental benchmark and concluded, in sharp contrast to LaLonde, that

> "the 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." (Dehejia and Wahba 1999, p. 1062).

The contrast between the conclusions in Dehejia and Wahba (1999) and those in LaLonde (1986) started an explosion of methodological work probing these conclusions. This led to the development of additional robust estimators, systematic methods for accounting for overlap, and placebo methods. We discuss these developments in the next section.

Methodological Improvements since LaLonde (1986)

This section begins by introducing the potential outcome framework. We then examine two key assumptions: unconfoundedness and overlap, followed by a brief discussion of various estimation strategies applicable under these assumptions. Next, we discuss alternative estimands, such as conditional average treatment effects and quantile treatment effects, and the methods for estimating them. Finally, we highlight the importance of supplementary analyses, primarily placebo tests, which are key to validating these key assumptions and improving research credibility.

Potential Outcome Framework

To facilitate the discussion of the modern causal inference literature, we adopt the potential outcome model originally used by Jerzey Neyman in the context of randomized experiments (Neyman 1923/1990), and extended to nonexperimental studies by Donald Rubin (Rubin 1974, 2006). For each individual i, for $i = 1, \ldots, N$, two potential outcomes

exist: $Y_i(0)$ represents the outcome (earnings in 1978) had individual i not participated in the NSW program, and $Y_i(1)$ represents the outcome for the same individual had they participated in the program. The difference between those two potential outcomes, $\tau_i \equiv Y_i(1) - Y_i(0)$ is the causal effect of the program for that individual. The binary treatment for individual i, participation in the job training program, is denoted by $W_i \in \{0, 1\}$. The realized outcome is $Y_i \equiv Y_i(W_i) = (1 - W_i)Y_i(0) + W_i Y_i(1)$. We also observe pretreatment characteristics for each individual. In the LaLonde case, the basic vector of covariates includes age, years of schooling, high school dropout status, marital status, and indicators for African-American and Hispanic backgrounds. Following Dehejia and Wahba (1999)'s analysis, we also consider settings where the covariate vector is augmented to include two lagged earnings variables—earnings in 1974 and 1975—and binary indicators for these lagged earnings being zero, indicating unemployment. X_i denotes the vector consisting of all ten pretreatment covariates.

Our primary estimand is the average treatment effect for the treated (ATT),

$$ATT \equiv \frac{1}{N_{tr}} \sum_{i:W_i=1} \left\{ Y_i(1) - Y_i(0) \right\},$$

where N_{tr} is the number of treated units. In other settings researchers may also be interested in the average treatment effect $ATE \equiv \frac{1}{N} \sum_{i=1}^{N} \left\{ Y_i(1) - Y_i(0) \right\}$. Most analyses of the LaLonde data that explicitly allow for treatment effect heterogeneity focus on the ATT, as it makes no sense to estimate or even contemplate the effect of the program for those individuals in the control group who have long term jobs and high earnings. LaLonde (1986) does not draw a distinction between these two estimands as he generally did not explicitly discuss effect heterogeneity.

Of course, we cannot directly estimate the ATT because we do not observe the control outcomes for the treated units, what (Holland 1986) called the "fundamental problem of causal inference." To make progress, let us define the *statistical estimand*, the covariate adjusted difference in the average outcomes between treated and controls,

$$\mathbb{E}\left[\mathbb{E}[Y_i | W_i = 1, X_i] - \mathbb{E}[Y_i | W_i = 0, X_i] \mid W_i = 1 \right].$$

This is an object we can estimate consistently given a random sample. However, it is only under two critical assumptions, unconfoundedness and overlap, that this statistical estimand is equal to the *causal estimand*, the ATT, which is the object of interest.

Part of the subsequent literature has focused on better statistical methods for estimating the statistical estimand, the covariate adjusted difference, increasingly relying on insights from the machine learning literature to deal with the potentially high-dimensional nature of the pre-treatment variables. Formal results typically require additional regularity conditions, such as the smoothness of conditional means and

propensity scores, along with moment conditions. For a more formal treatment of this topic, we refer readers to the original papers, as referenced in reviews such as Imbens and Wooldridge (2009) and Abadie and Cattaneo (2018). It should be noted that this is a purely statistical issue, and one that does not depend on whether unconfoundedness holds, that is, on whether the statistical estimand is actually of interest.

A separate question concerns the plausibility of the assumptions, and in particular the unconfoundedness assumption. Without that assumption, we may be able to estimate the covariate-adjusted difference robustly, but it may not be of any interest. Regarding this second question, the literature has also made substantial progress. First, through recognizing that this is in fact a separate issue, and second through the development of placebo and sensitivity analyses.

Unconfoundedness

The unconfoundedness assumption, first introduced by Rosenbaum and Rubin (1983b) as part of the "ignorable treatment assignment" concept that also includes overlap, has played a crucial role in identifying average treatment effects using nonexperimental data. It states that, conditional on the covariates, the treatment assignment is independent of the pair of potential outcomes:

Assumption 1 (Unconfoundedness).

$$W_i \perp\!\!\!\perp \{(Y_i(0), Y_i(1)\} \mid X_i.$$

Identifying the ATT in fact only requires $W_i \perp\!\!\!\perp Y_i(0) \mid X_i$, a weaker version of unconfoundedness. The unconfoundedness assumption is also referred to as *exogeneity* (Imbens 2004), *conditional independence* (Lechner 1999, 2002) or *selection on observables* (Barnow et al. 1980). This assumption stands in contrast to traditional econometric definitions of exogeneity that were articulated in terms of residuals, themselves defined in terms of functional forms. Unconfoundedness elegantly separates the functional form part of the assumptions from their essence. Essentially, it is sufficient that researchers understand (a crucial aspect of) the *design*, or the treatment assignment mechanism, without full knowledge of the data-generating process of the potential outcomes. A key result in Rosenbaum and Rubin (1983b) shows that Assumption 1 implies

$$W_i \perp\!\!\!\perp \{(Y_i(0), Y_i(1)\} \mid e(X_i),$$

in which $e(X_i) \equiv \Pr(W_i = 1 \mid X_i)$ is the propensity score for unit i. This result is important in guiding many estimation strategies because it reduces the dimension of the conditioning set from the number of pre-treatment variables (which can be substantial) to one, the dimension of the propensity score.

When the parametric outcome model (*e.g.*, the earnings equation) is correctly specified, unconfoundedness implies a zero conditional mean for the error term. Thus, except for those using a difference-in-differences approach, the estimates in Tables 4 and 5 of LaLonde (1986) can be interpreted as based on a combination of unconfoundedness and functional form assumptions. At the time of LaLonde's study, however, the design-based perspective had not yet gained popularity, and these specifications were motivated almost entirely from an outcome modeling perspective.

In practice, unconfoundedness is a very strong assumption. For a general discussion on this topic, we recommend Rosenbaum and Rubin (1983b) and Imbens (2004). While we acknowledge concerns about its validity in the absence of a clear understanding of the treatment assignment mechanism, we believe that supplementary analyses using ancillary data and domain knowledge, such as placebo tests and sensitivity analyses, can help assess the plausibility of this assumption and, by doing so, improve the credibility of analyses based on it. We will illustrate the usage of placebo tests in the next section. In the context of the LaLonde data, it is evident that all ten covariates are appropriate pretreatment variables that should be controlled for. In other cases, whether one should adjust for differences between treated and control units based on specific covariates is less clear. Rosenbaum (1984) cautions against adjusting for variables that are affected by the treatment. Cinelli et al. (2022) further discuss the selection of variables within the set of proper pre-treatment variables to adjust for in causal analyses.

Overlap and Balance

To identify the average causal effect under unconfoundedness, we need to ensure that we can estimate the average effect at every value for the covariates, requiring *overlap*, or that the propensity score is between zero and one:

Assumption 2 (Overlap).

$$0 < \Pr(W_i = 1 \mid X_i) < 1.$$

If the ATT is of interest, in fact only a weaker overlap assumption, $\Pr(W_i = 1 \mid X_i) < 1$, is required. Overlap is crucial in identifying the ATT when researchers are unwilling to make functional form assumptions about the conditional means of the potential outcomes and the extent of heterogeneity in the treatment effects. When X_i includes fewer than a handful of covariates, inspecting pairs of the covariates' marginal or joint distributions by treatment status may be sufficient for assessing overlap. However, this approach becomes impractical in high-dimensional settings. In such cases, a more attractive method is to inspect the distribution of the propensity scores, estimated by a flexible method, by treatment status. The lack of overlap in covariate distributions implies, and is implied by, a lack of overlap in the propensity score distributions.

LaLonde (1986) did not explicitly discuss overlap, nor did he inspect it beyond report sample averages for covariates by treatment status. Both the regression and selection model approaches he used assume correct functional forms, which allow for interpolation or extrapolation of treatment effects across all covariate levels and their combinations, thereby formally eliminating the need for overlap. However, even in the absence of formal methods for addressing the full extent of the overlap assumption, LaLonde was clearly concerned about the possible implications of lack of overlap between the experimental treated group and the comparison group based on CPS and PSID data. As mentioned earlier, to improve comparability, he trimmed the original 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), aligned with the goal of improving balance on covariates that determine selection. However, by modern standards, his methods, such as removing all male participants working in March 1976 (CPS1-SSA-2) or further removing unemployed respondents with 1975 incomes above the poverty line (CPS1-SSA-3), are *ad hoc* and do not necessarily achieve overlap in all relevant covariates.

Since LaLonde (1986), many scholars have proposed more principled and systematic methods to improve overlap, often relying on propensity scores. These methods take different forms, partly depending on whether simply overlap in the covariate distributions is sought, or whether, more aggressively, balance in the covariate distributions is pursued. Overlap refers to the difference in the range of covariate values in treatment and control groups. Balance refers to the similarity of the covariate distributions in treatment and control groups. In expectation, balance is achieved by design in a completely randomized experiment, and can be further improved upon through stratification prior to the randomization.

Ensuring overlap or improving balance typically involves dropping some units from the full sample. Although in principle this leads to some loss of information, the improvement in robustness and reduction of bias may outweigh the loss in precision. In fact, the potential increase in variance from a substantial amount of trimming of the sample is typically modest. Suppose one has a sample with N_{tr} treated units, and N_{co} control units. Under homoskedasticity and random assignment, the variance of the difference in mean estimator is $\sigma^2(1/(N_{tr} + 1/N_{co})$. For instance, if we start with $N_{tr} = 185$ treated units and $N_{co} = 15,922$ control units as in the LDW-CPS sample, dropping 15,737 control individuals to leave just 185 control individuals (a 99% reduction in the control sample) increases the standard error only by 30% in the "best-case scenario," which assumes no bias from including the additional control individuals. In practice, concerns about bias suggest that aggressive trimming may lead to more robust and credible estimates.

Focusing on overlap alone, and with the estimand the ATT, Dehejia and Wahba (1999)

drop all control individuals with a propensity score less than the smallest propensity score among the treated individuals. Crump et al. (2009) develop a more aggressive approach to address the lack of overlap. They characterize subsamples optimized for precise average treatment effect estimation, with a rule of thumb suggesting trimming data with estimated propensity scores outside $[0.1, 0.9]$. Crump et al. (2006) and Li et al. (2018) propose balancing covariates through propensity score weighting, introducing the "overlap weights" that are proportional to the product of the propensity score and one minus the propensity score. A third approach, particularly well suited to settings where the focus is on the ATT, is to create a matched sample in which all treated units are matched to a distinct control unit in terms of the estimated propensity score. Beyond ensuring overlap, this method creates a sample that is much better balanced in the covariate distributions.

In practice, overlap, like unconfoundedness, is critical for obtaining credible estimates. This is particularly true in cases with poor overlap in the raw data, such as the LaLonde samples. For such settings, trimming to ensure overlap is often more important than the choice of specific estimation strategies.

Estimation Given Unconfoundedness and Overlap

All estimators in LaLonde (1986) are linear in the covariates. Subsequently, a variety of methods have been proposed to estimate average causal effects in more flexible ways under both unconfoundedness and overlap assumptions, a combination also referred to as ignorable treatment assignment (Rosenbaum and Rubin 1983b). We divide these methods into three groups: (i) outcome modeling, including linear regressions, (ii) methods that directly adjust for covariate imbalance, including those based on propensity scores, and (iii) doubly robust methods.

Outcome modeling. The simplest and still the most commonly used method by applied researchers is a simple linear regression using the treatment indicator and covariates (the level terms) as regressors, which resembles the earnings equation in LaLonde (1986). The regression method models the conditional means of potential outcomes parametrically and requires the treatment effect to be constant. Relaxing the functional form assumptions slightly, one can use two separate linear regressions to model the conditional means of the two potential outcome. This estimator is sometimes referred to as the Oaxaca-Blinder estimators (Kline 2011). More generally, researchers can model the two conditional means using semiparametric or nonparametric approaches (*e.g.*, Heckman et al. 1997, 1998; Athey et al. 2019).

Adjusting covariate imbalance. The second group of methods focuses on directly adjusting covariate imbalance between the treatment and control groups. This includes blocking on covariates, covariate matching (e.g., Abadie and Imbens 2006, 2008, 2011, 2016; Diamond and Sekhon 2013; Imbens 2015), and weighting methods to achieve covariate balance (*e.g.*, Hirano et al. 2003; Hainmueller 2012; Zubizarreta et al. 2023; Zubizarreta 2015).

When the number of covariates is large, particularly with many continuous variables, simple covariate matching methods suffer from the curse of dimensionality, rendering them either infeasible or prone to large biases (Abadie and Imbens 2006). Under such circumstances, adjusting for differences in the propensity score, rather than attempting to adjust for all covariates, is often beneficial. This can be implemented through various methods, such as blocking/matching (*e.g.*, Dehejia and Wahba 1999) and inverse propensity score weighting (IPW). For example, Hirano et al. (2003) show that this Hájek estimator, a variant of the IPW estimator, can achieve the semiparametric efficiency bound with a nonparametric estimator for the propensity score.

There are also attempts to improve covariate balance while estimating the propensity score or without directly estimating it. For example, Imai and Ratkovic (2014) propose covariate balancing propensity score estimated via the generalized method of moments using covariate balance as moment conditions. Hainmueller (2012) proposes entropy balancing to directly adjust for covariate imbalance. Research shows that entropy balancing can be seen as an IPW estimator with a linear propensity score model and a logistic link (Zhao and Percival 2016).

Doubly robust methods. Neither outcome modeling nor the balancing methods are currently the most recommended methods in the methodological literature. Instead, scholars have developed various mixed methods that combine outcome modeling, such as regression, with methods addressing covariate imbalance to achieve the benefits from both. These methods include regression within propensity score blocks (Rosenbaum and Rubin 1983a; Imbens 2015), matching combined with regression (Abadie and Imbens 2011), and methods integrating weighting with regression (e.g., Robins et al. 1994; Robins and Rotnitzky 1995). The rationale for the mixed methods is that, although covariate-balancing or propensity score methods by themselves may be consistent, or even fully semiparametrically efficient, incorporating outcome models can improve small sample performance by eliminating remaining biases or improving precision by leveraging the correlation between the covariates and the outcome. For instance, while the bias of a simple matching estimator might dominate variance in high-dimensional cases, adding regression to account for the remaining imbalance can substantially reduce such biases (Abadie and Imbens 2011).

Robins and Ritov (1997) introduce the term "double robustness," an important concept for these mixed methods. They show that if either the propensity score or regression model is correctly specified parametrically, the augmented inverse propensity weighting (AIPW) estimator that combines weighting and regression is consistent. They can be viewed as combining an outcome model with an adjustment term, which consists of an IPW estimator applied to the residuals from the outcome model.

In the past few years, machine learning methods have rapidly entered the toolkit of applied researchers for estimating causal effects due to advancement in the methodological literature (Van der Laan and Rose 2011; Chernozhukov et al. 2017; Athey et al. 2018, 2019). Many of these estimators adopt the form of an AIPW estimator and satisfy the "Neyman orthogonality" condition (Chernozhukov et al. 2018), which ensures the stability of the moment conditions used to identify the causal parameter against small perturbations in nuisance functions, including the conditional mean and the propensity score. Chernozhukov et al. (2017, 2018) show a particularly attractive feature of what they labeled the double/debiased machine learning estimators based on estimating the influence function for the semiparametrically efficient estimator: they accommodate slower convergence rates for the estimators of the nuisance functions.

Alternative Estimands and Heterogeneous Treatment Effects

Much of the methodological and applied research has focused on estimating average treatment effects, such as the ATT. However, there are other quantities of interest to researchers. For example, researchers are often interested in understanding variation in the treatment effects. Understanding effect heterogeneity is crucial for discerning the mechanisms and impacts of a policy, for a more precise evaluation of policy effectiveness, and for guiding personalized policy assignments. Econometrically, researchers can study heterogeneous treatment effects by estimating the conditional average treatment effect on the treated (CATT), i.e., $\tau(x) \equiv \frac{1}{N_x} \sum_{i:X_i=x,W_i=1} \tau_i$, in which N_x is the number of treated units whose covariate values equal to x. Researchers have proposed to use machine learning methods to estimate CATT nonparametrically or using low-dimensional representations, such as causal forests, and obtain valid inference or error bounds (*e.g.*, Athey and Imbens 2016; Wager and Athey 2015; Athey et al. 2019).

Another important but less commonly used group of estimands by empirical researchers are the quantile treatment effects. They are defined as the difference between the quantiles of the treated and untreated potential outcome distributions for the population or the treated group. Because Assumptions 1 and 2 allow for the identification of the full marginal distribution of $Y_i(0)$ and $Y_i(1)$, quantile treatment effects are identified under those assumptions. Firpo (2007) proposes a semiparametrically efficient estimator for these quantities, combining conditional quantile estimations with IPW. See Bitler

et al. (2006) for an application to the LaLonde data. One potentially underexplored area in the literature is that estimates of CATT or quantile treatment effects can inform the plausibility of the unconfoundedness assumption, given that researchers often possess insights into the range of these effects.

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 approaches: placebo analyses and sensitivity analyses. Due to space limitations, we discuss the former and relegate the latter to the online appendix.

Placebo analyses indirectly assess unconfoundedness by formally testing a conditional independence restriction. This testable assumption differs from unconfoundedness in two aspects. First, it conditions on a subset of the full set of covariates that appear in the unconfoundedness assumption. Second, it uses one of the remaining covariates as a pseudo-outcome that serves as a proxy for the target outcome. A common placebo test estimates the treatment effect on a pretreatment variable, known to be unaffected by the treatment. A lagged outcome is an appealing choice as it is typically a good proxy for the target outcome.[7]

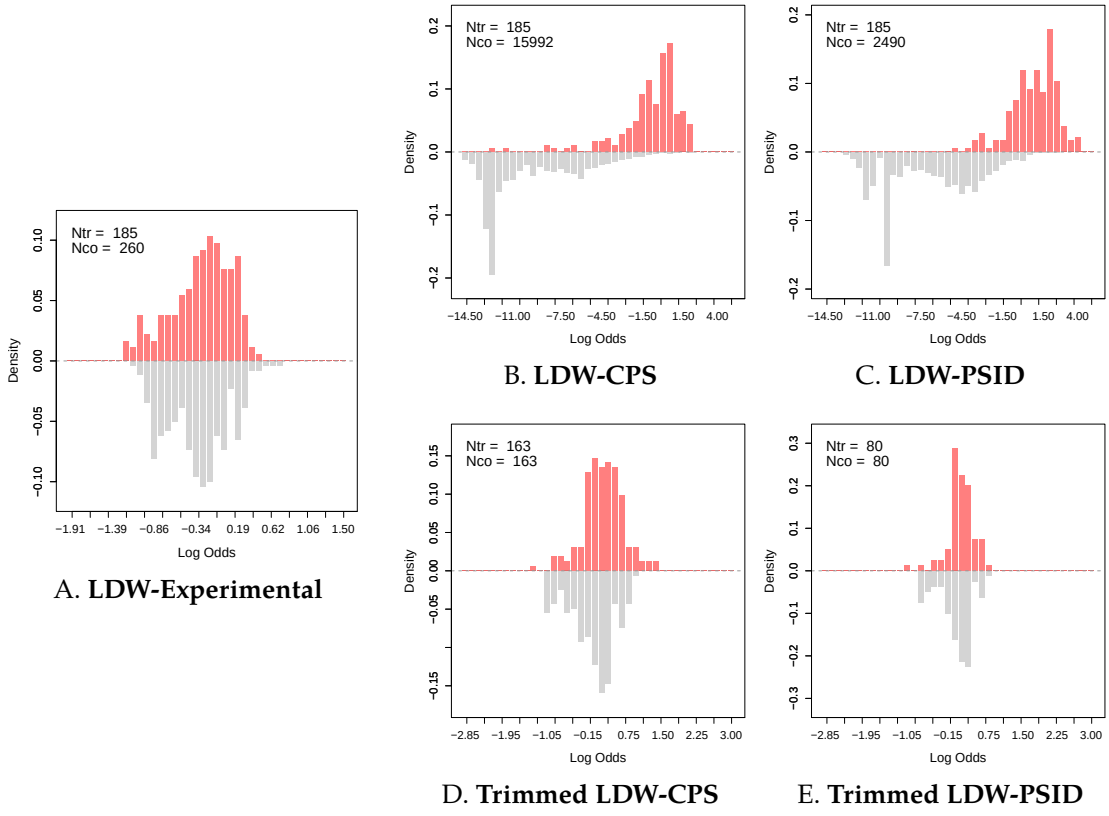
Lalonde regressed 1975 earnings, which predated the program, on the treatment indicator and covariates, and reported findings in columns 2 and 3 of Tables 4 and 5. He found that most nonexperimental estimates are negative, large, and often statistically significant, indicating a potential violation of unconfoundedness. Although LaLonde did not explicitly use the term, this approach is what we would now call a placebo test. One limitation of the LaLonde data is the availability of only one pretreatment outcome. With the Lalonde-Dehejia-Wahba data, we can test whether 1975 earnings are correlated with participation in the job training program, conditional on 1974 earnings and other covariates. With additional pretreatment periods, as in the Imbens-Rubin-Sacerdote data (with six lagged outcomes), researchers can construct placebo tests that are statistically more powerful and substantively more credible. A limitation of LaLonde's analyses is that he only tested one aspect of the full conditional independence assumption, *i.e.*, whether the two conditional means, averaged over the conditioning variables, are the same. Imbens (2015) discusses testing additional implications of the conditional independence relationship.

[7]Other forms of placebo tests include estimating the effect of a pseudo-treatment on the outcome, often using multiple control groups. For further details, see Rosenbaum et al. (1987); Imbens and Rubin (2015); Imbens (2015).

Reanalyzing the LaLonde and Imbens-Rubin-Sacerdote Data

To demonstrate the methodological advances since LaLonde (1986) in practice, we revisit the LaLonde data, including both the LaLonde-Dehejia-Wahba data, the original Lalonde male samples, and the Lalonde-Calónico-Smith female samples. Overlap is a major concern for all of these datasets. We also analyze the Imbens-Rubin-Sacerdote lottery data, where extensive pretreatment information is available and overlap is less concerning. We focus on the ATT as the main causal quantity of interest for both the LaLonde and Imbens-Rubin-Sacerdote datasets.

FIGURE 1. **Assessing the Overlap in Lalonde-Dehejia-Wahba (LDW) Data**



Note: Histograms depict the log odds ratios, i.e., $\log \frac{\hat{e}}{1-\hat{e}}$, using propensity score estimated through generalized random forest. Each subfigure represents a different sample. N_{tr} and N_{co} represent the numbers of treated and control units, respectively. **Subfigure A**: LDW-Experimental. **Subfigure B**: LDW-CPS. **Subfigure C**: LDW-PSID. **Subfigure D**: Trimmed LDW-CPS. **Subfigure E**: Trimmed LDW-PSID. For C and D, the propensity scores are re-estimated after trimming.

The LaLonde Data

We primarily focus on the Lalonde-Dehejia-Wahba (LDW) data because information on earnings and employment status in 1974 is available. We use three LDW datasets: (1) LDW-Experimental, which consists of 185 treated and 280 control individuals from the

experimental data; (2) LDW-CPS, including the same treated individuals and 15,992 controls from CPS-SSA-1; and (3) LDW-PSID, comprising the same treated individuals and 2,490 controls from PSID-1. We do not use CPS-SSA-2 and CPS-SSA-3 controls as they are subsets of CPS-SSA-1, and similarly, we do not use PSID-2 and PSID-3 as they are part of PSID-1. LaLonde's construction of these subsamples was a relatively *ad hoc* approach to improving overlap. We use more modern, fully data-driven ways to detect and address issues related to overlap.

Figure 1 (A)-(C) display the overlaps in propensity score estimated from Generalized Random Forest (GRF, Athey et al. 2019) between the treated and control units for all three samples, using histograms of the log-odds of propensity scores, *i.e.*, $\log(\hat{e}/(1-\hat{e}))$. In (A), as expected, LDW-Experimental shows almost perfect overlap, with the distributions of estimated propensity scores closely mirroring each other between the treatment and control groups. In (B) and (C), however, both nonexperimental samples show very poor overlap and balance. Most notably, the propensity scores of many treated units do not lie within the support of the controls' propensity scores, and a substantial proportion of the control units possess extremely low log odds. Similar patterns are observed with the original LaLonde male samples presented in the online appendix.

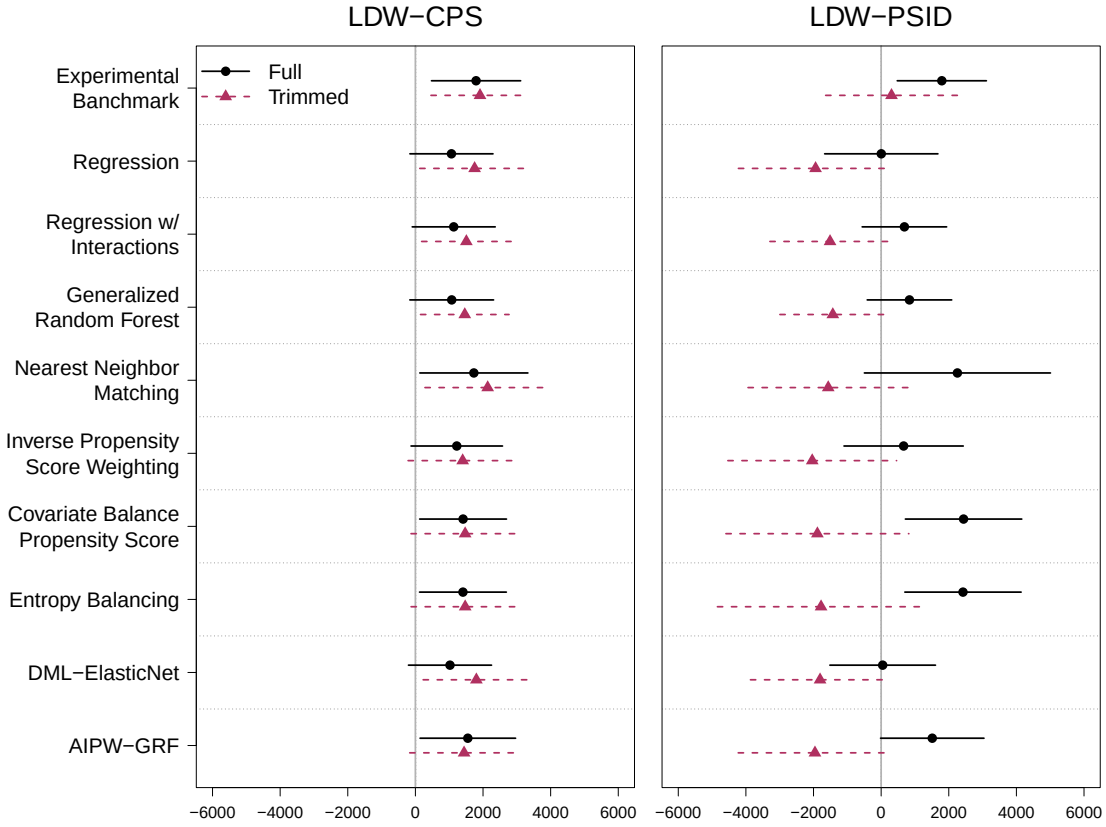
Trimming to improve overlap. We construct two trimmed samples using LDW-CPS and LDW-PSID to improve overlap between experimental and nonexperimental units, which takes two steps. We start by merging experimental controls from LDW-Experimental into LDW-CPS and LDW-PSID and estimating the propensity of each unit being included in the experiment using GRF. We then trim the based on set thresholds, resulting in the exclusion of some treated units (Crump et al. 2009). After trimming, we re-estimate the propensity scores using the remaining data, and perform 1:1 matching to further trim the nonexperimental controls. This procedure yields two sets of trimmed samples: one composed of experimental treated units and nonexperimental controls, and another serving as an experimental benchmark.[8] As shown in Figure 1(D)-(E), overlap improves significantly in both samples post-trimming, though this comes with the cost of reduced sample sizes.

Estimating the ATT. Next, we estimate the ATT using both the original Lalonde-Dehejia-Wahba nonexperimental samples and the newly constructed trimmed samples. We apply a variety of estimators, including simple difference-in-means, regression, regression with interactions (Oaxaca-Blinder), generalized random forest as an outcome model,

[8]The procedure is designed to improve overlap while obtaining experimental benchmarks for the trimmed samples. We discuss its details in the online appendix. When researchers have only one observational dataset, a one-step trimming based on the estimated propensity score is advised.

nearest neighbor matching with bias correction, IPW with propensity scores estimated by GRF, covariate balancing propensity score, entropy balancing, double/debiased matching learning using elastic net (DML-ElasticNet), and AIPW implemented via generalized random forest (AIPW-GRF). All estimators use the same set of ten covariates as before.

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



Note: The figures above show the ATT estimates and their 95% 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 (GRF) as an outcome model, 1: 5 nearest neighbor matching with bias correction, inverse propensity score weighting with propensity scores estimated by GRF, covariate-balance propensity score, entropy balancing, double/debiased machine learning with elastic net (DML-ElasticNet), implemented using `DoubleML`, and augmented inverse propensity score weighting (AIPW) with GRF, implemented using `grf`. Difference-in-means estimates are not shown because they are extreme with LDW-CPS and LDW-PSID at $-8,497 and $-15,204, respectively, and similar to other estimates in the two trimmed samples at $1,483 and $-1,505.

We present the findings in Figure 2. The first two panels present the ATT estimates and their 95% confidence intervals from LDW-CPS and LDW-PSID, while the third and fourth panels show the results from the trimmed samples with improved overlap. In each figure, the ATT estimates using experimental data and their 95% confidence

intervals are highlighted with a red dashed line and pink band, respectively.

As shown in the first column of Figure 2, when using LDW-CPS, all estimators, except difference-in-means, produce positive estimates, although there are noticeable variations among them. Nearest neighbor matching outperforms other estimators, aligning closely with the experimental benchmark of $1,794. Notably, covariate balancing propensity score, entropy balancing, and AIPW-GRF also produce results close to the benchmark. Despite numerical differences, these estimates, except for difference-in-means, cannot be statistically distinguished from one another. The second column of Figure 2 shows that estimates based on LDW-PSID exhibit greater variations. Setting aside the difference-in-means, the estimates span from $4 to $2,420. Among them, the AIPW-GRF estimator produces an estimate closest to the experimental benchmark.

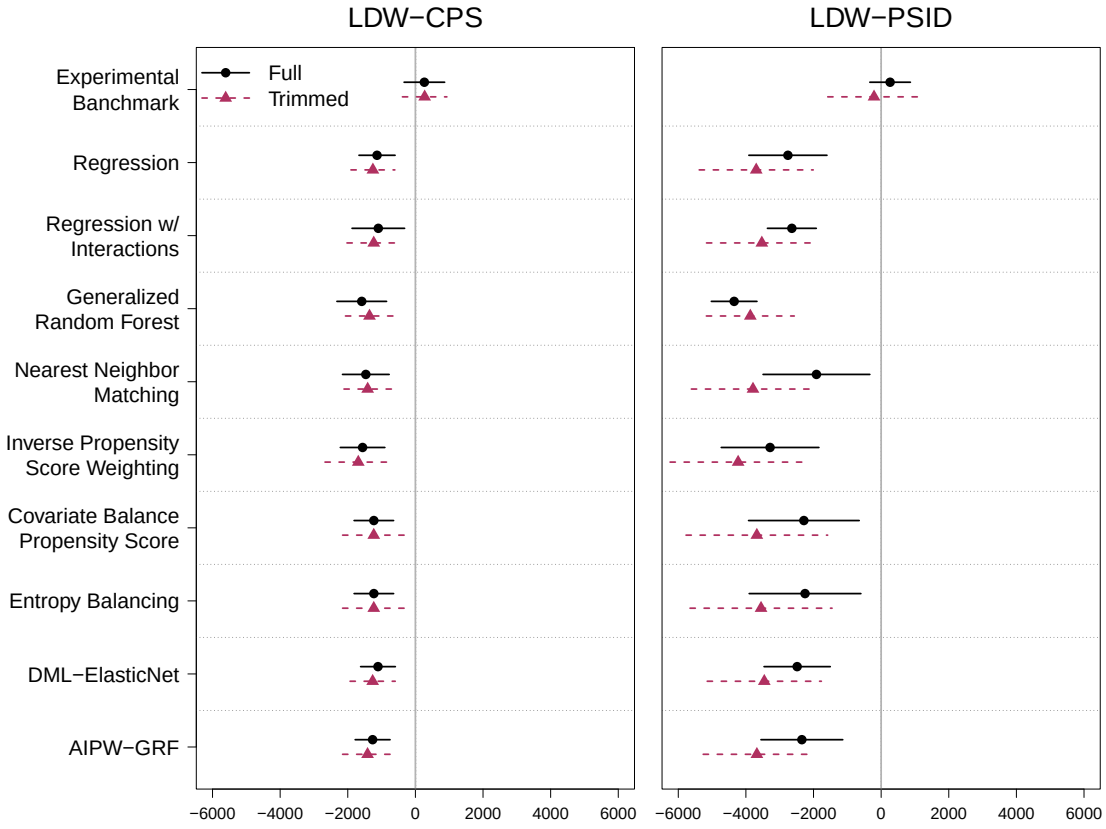
The last two columns of Figure 2 show that by using trimmed data with improved overlap, estimates produced by various estimators are substantially more stable. For trimmed LDW-CPS, all estimates hover around the experimental benchmark of $1,911. In the case of trimmed LDW-PSID, the experimental benchmark stands at $306, and it is not statistically significantly different from zero at the 5% level. While each estimator yields results that are closely aligned, they all have a negative sign. They are statistically indistinguishable from the experimental benchmark due to the large uncertainties associated with these estimates. The stability of the range of estimators after trimming is also found in the simulations in Athey et al. (2021).

These findings suggest that improved overlap based on observed covariates can reduce model dependency and estimate variability across different estimators, leading to more robust estimates of the statistical estimand. However, this does not guarantee consistency without validating unconfoundedness. The fact that many methods produce estimates that match the experimental benchmark for the ATT using LDW-CPS might have instilled unwarranted confidence in researchers that modern estimators can help achieve causal identification even when there are no compelling reasons to believe unconfoundedness. In other words, while modern methods may be effective in estimating the statistical estimand, the covariate-adjusted difference in the average outcomes between the treated and control groups, this does *not* mean that the adjusted difference is close to the causal estimand, which is the ATT. For that, we need some version of the unconfoundedness assumption, which is fundamentally untestable. However, we can assess its plausibility, even if we cannot formally test it.

Validation through placebo analyses. We conduct placebo analyses to further assess how plausible unconfoundedness is. To do so, we select earnings in 1975 as the placebo outcome and remove both earnings in 1975 and employment status in 1975 from the set of conditioning variables. Two new trimmed samples are also created without using earnings and employment status in 1975. We then estimate the ATT for the placebo

outcome, adjusting for the remaining covariates using a variety of estimators. Figure 3 presents the findings. Not surprisingly, the experimental benchmarks are near zero and statistically insignificant. However, all estimators using nonexperimental data generate large, negative estimates. Again, with trimmed data, the estimates are stable but remain statistically different from zero. Moreover, we further show in the online appendix that that while the CATT estimates from the experimental data hover close to zero, their nonexperimental counterparts are all negative and substantial in magnitude, indicating large biases.

FIGURE 3.
Placebo Tests: '75 Earnings as the Outcome



Note: The figures above show the placebo estimates and their 95% 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. We use the same ten estimators as before. The difference-in-means estimates are not shown; they are $-12,118, $-17,531, $-14,56, and $-4,670 in the four panels, respectively.

Alternative samples. For comparison, we also revisit the original male samples used in LaLonde (1986) and AFDC female samples reconstructed by Calónico and Smith (2017). Information on 1974 earnings and employment status is unavailable in these datasets. We report the findings in the online appendix. For the LaLonde male sample, we find that, with sufficient overlap, most modern estimators yield estimates within

relatively narrow ranges when using either CPS-SSA-1 or PSID-1 as control groups. However, these estimates do not align with the experimental benchmarks, with most estimates being negative. Smith and Todd (2001, 2005) report similar, negative findings.

Using the Lalonde-Calónico-Smith (LCS) female samples, we find that many modern methods yield estimates close to the experimental benchmarks, though standard errors are often quite large. While selection appears to be less severe for AFDC women compared to the male NSW participants, as suggested by Calónico and Smith (2017), overlap remains a significant challenge. Additionally, we fail to substantiate the unconfoundedness assumption with a placebo test using the number of children in 1975, a variable absent in LaLonde's anlysis, as the placebo outcome.

Summary. After reexamining the LaLonde data, we offer some new insights into the challenge posed by LaLonde. First, we agree with existing literature that ensuring overlap and using comparable control units are essential for credible causal estimates. Second, while the choice of method is less critical with overlap, as most methods yield similar results, the propensity score remains a vital tool for assessing overlap and is integral to many estimators. Moreover, we stress the need for additional tests to validate unconfoundedness. With the LDW and LCS data, many methods approximate the experimental benchmark for the *average effects* under overlap, a success not mirrored with the original LaLonde male samples. However, even with LDW or LCS data, placebo tests fail to support unconfoundedness.

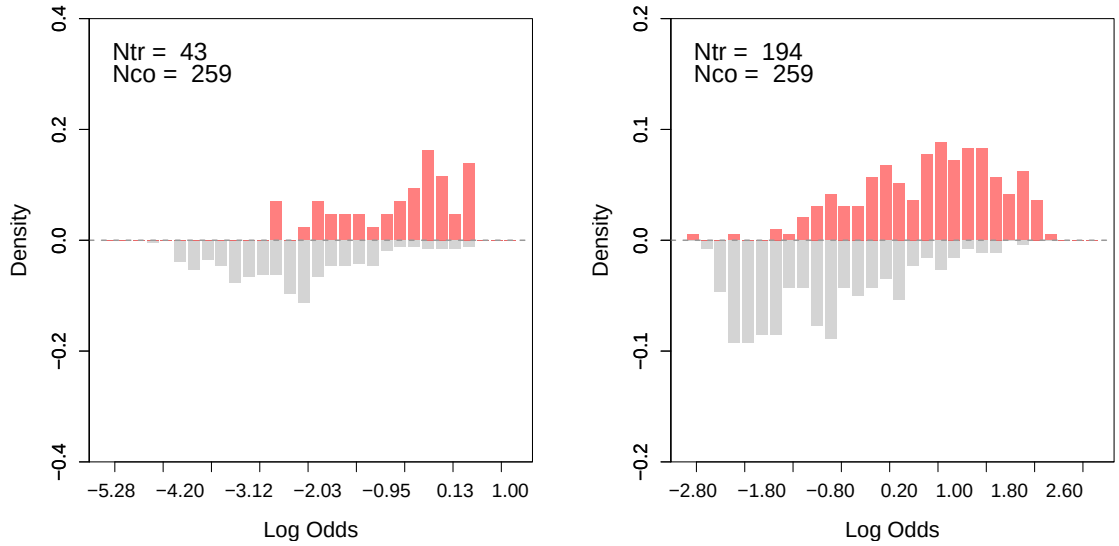
Lottery Prizes on Labor Earnings

We now turn to the Imbens-Rubin-Sacerdote lottery data. The authors carried out an original survey to investigate the impact of the size of lottery prizes in Massachusetts during the mid-1980s on the economic behavior of lottery players. The primary outcome is post-winning labor earnings. This empirical example is appealing for two reasons: (i) we have a much better understanding of the treatment assignment process (lottery), and (ii) six periods of lagged outcomes are available to validate the unconfoundedness assumption.

There are three treatment and control groups. The control group, termed "non-winners," consists of 259 season ticket holders who have won a small, one-time prize, ranging from $100 to $5,000 (in essence, they are one-time, minor winners). The treatment groups, labeled "big winners" (43 individuals) and "small winners" (194 individuals), are those who clinched a major prize. They might be season ticket holders or one-time buyers. The annual installments for these prizes ranged from $1,139 to $99,888 (small winners) and exceeded $100,000 (big winners), respectively. These prizes were disbursed in yearly installments for over 20 years.

While randomization should ideally ensure that the treatment and control groups are comparable at the time of the lottery entry, the authors highlight three potential reasons this might not be the case. First, individuals can purchase multiple tickets, increasing their odds of winning. Second, those who hold season tickets might differ from those who buy single tickets. Lastly, there were discrepancies in the response rates between winners and non-winners (49% and 42%, respectively), and these response rates could be influenced by a range of factors, as evidenced by the decline in response probability with the magnitude of the prize. However, the authors expect that the unconfoundedness assumption will hold once they condition on a set of observable covariates, including the year of winning and the number of tickets bought. Importantly, they also gathered data on past labor earnings for up to six years before the individuals won a prize. These past outcomes can be utilized either as conditioning variables or as placebo outcomes.

FIGURE 4.
Assessing Overlap in the Imbens-Rubin-Sacerdote Lottery Data



A. **Big Winners vs Non-Winners** B. **Small Winners vs Non-Winners**

Note: Histograms depict the log odds ratios, i.e., $\log \frac{\hat{e}}{1-\hat{e}}$, using propensity scores estimated through generalized random forest. N_{tr} and N_{co} represent the numbers of treated and control units, respectively.

In the subsequent analysis, we will consider labor earnings from seven post-lottery-winning periods as the outcomes. These are denoted as $Y_{i,0}, \ldots, Y_{i,6}$, where $t = 0$ represents the year of winning a lottery—recall that individuals in the control group also received a modest, one-time prize that year. We will treat the labor earnings from the three years immediately preceding the lottery win, i.e., $Y_{i,-3}$, $Y_{i,-2}$, $Y_{i,-1}$, as well as their average, as placebo outcomes. The labor earnings from the three years before those, i.e., $Y_{i,-6}$, $Y_{i,-5}$, $Y_{i,-4}$, will be used as covariates for adjustment, alongside a set of time-invariant pre-lottery-winning variables. These include the number of tickets purchased, gender, employment status at the time of winning, age when the lottery was won, total

years of education, and the presence of a college degree. Figure 4 assesses the overlap between the two treatment groups and the control group using the mentioned covariates. The figure indicates that while the propensity score distribution of individuals in the treatment groups differ from that of the control group, the propensity scores of the treatment groups still fall within the support of the control group.[9]

We estimate the ATT for labor income from Year −3 to Year 6 separately using both difference-in-means and AIPW-GRF. Figure 5 shows the results. The representation resembles an event study plot used in panel data analyses, although our main identification assumption is unconfoundedness. In estimating the effect of big prizes, AIPW-GRF using the original or trimmed data produces estimates very similar to a simple difference-in-means estimator, suggesting minimal selection between the two groups. On the other hand, when estimating the effect of small prizes, the estimates from AIPW-GRF and difference-in-means diverge. However, findings from the former are much more credible than those from the latter because difference-in-means does not fare well in the placebo tests, whereas the former yields placebo estimates that are nearly zero. AIPW-GRF using either the original or the trimmed sample produce results aligned with the findings reported in the original paper: winning a large prize leads to a significant decrease in labor income in the following years, averaging as much as $8,000 annually. In contrast, winning a smaller prize results in a more modest decline, averaging approximately $3,000 per year.

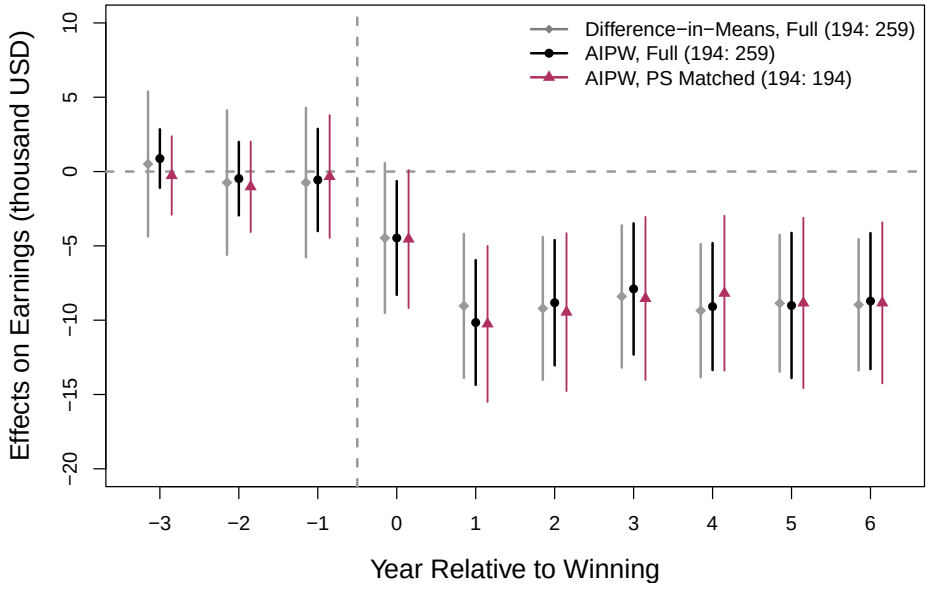
In the lottery study, placebo tests provide strong evidence for the unconfoundedness assumption, bolstering the credibility of the causal estimates. Importantly, unconfoundedness is much more believable in this study than in the LaLonde case because the inherent randomization of lotteries played a key role in treatment assignment, while supplementary covariates help account for discrepancies between treatment and control groups stemming from challenges like differential responses to the survey. The inclusion of six preceding outcomes also proves invaluable, as they likely explain both selection and the outcome variables; moreover, they also serve as good candidates for placebo outcomes, given their comparability to these outcomes.

Lessons Learned

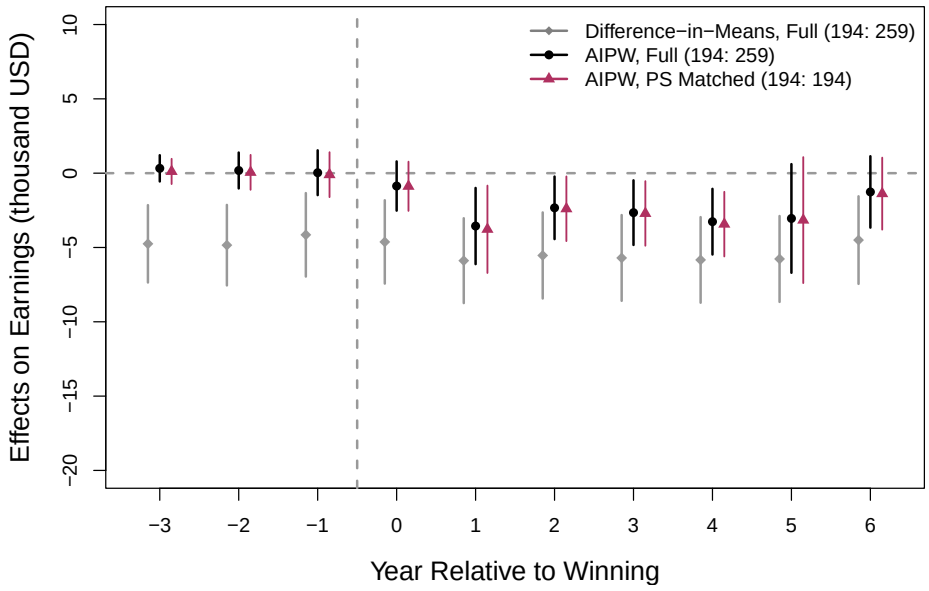
What specifically has the methodological literature since LaLonde (1986) taught us? What particular analyses would we recommend a researcher analyzing data of this type do in the light of the subsequent theoretical research? What insights do the reanalyses of the LaLonde and Imbens-Rubin-Sacerdote datasets provide?

[9]To improve overlap, we further trim the control group for each of the two treatment groups by implementing 1:1 matching based on propensity scores, resulting in two trimmed samples.

FIGURE 5.
Lottery Prizes on Labor Earnings: Imbens-Rubin-Sacerdote Data



A. **ATT: Big Winners vs Non-Winners**



B. **ATT: Small Winners vs Non-Winners**

Note: Figures show the ATT estimates using the Imbens-Rubin-Sacerdote data. The outcome variables include earnings from 3 years before winning to 6 years after winning. The estimates for pre-winning outcomes serve as placebo tests. Adjusted covariates include: time of playing, #tickets bought, gender, work then, age at winning, years of education, college degree, and earnings 6 to 4 years before winning. We use the difference-in-means estimator (gray diamonds) and the AIPW-GRF estimator (black solid circles for the original data and red triangles for the trimmed data).

First, for the type of data used in LaLonde's paper, the literature has predominantly focused on methods based on the unconfoundedness (or, using other terms, exogeneity, ignorability, conditional independence, or selection-on-observables) assumption.

Although LaLonde (1986), as well as Heckman and Hotz (1989) and others, explored alternative identification strategies in their analyses, including difference-in-differences and selection models, these methods are rarely applied to the LaLonde data in the subsequent literature. No compelling case has been made that other identification strategies are credible in that context. The lack of alternatives, of course, does not make the unconfoundedness assumption itself credible. The positive case for an analysis based on unconfoundedness is as follows: with the LaLonde data, or similar datasets, we argue that comparing treated and control untis identical or similar in terms of the full set of available pretreatment variables makes a causal interpretation more plausible than any other comparison between treated and control units. Any alternative strategy that would lead to point-identification would involve comparing treated and control units with different values for the pretreatment variables, which, in our view, makes a causal interpretation less credible. For that reason, we focus in this discussion on methods relying on unconfoundedness assumptions. For recent reviews of panel methods, see Xu (2023) and Arkhangelsky and Imbens (2024).

Second, and this is perhaps the most important insight, the literature has recognized the crucial role of assessing overlap in covariate distributions and dealing directly with the lack thereof. The various comparison groups LaLonde (1986) used to evaluate nonexperimental methods all differ substantially from the experimental sample in terms of the distributions of the covariates. This creates challenges for conventional statistical adjustment methods such as regression and matching. LaLonde attempted to address these by simply discarding individuals in the comparison groups who do not meet certain specific eligibility criteria based on age, employment status and earnings. The subsequent literature has emphasized that, in practice, lack of overlap is a key issue in such analyses. Effective and systematic, data-driven, ways of diagnosing, and addressing this lack of overlap have been developed subsequently.

Third, and somewhat related to the overlap issue, the role of the propensity score in estimation has been stressed. There are two components to this role. First, the propensity score plays an important role in uncovering and addressing the lack of overlap in covariate distributions. Second, it is important in the estimation of treatment effects, either directly through inverse propensity score weighting (IPW) or, more importantly in the current state-of-the-art approaches, as part of doubly robust methods that incorporate both models for the conditional outcome distributions and the models for the propensity score. Propensity scores played no role in LaLonde's analyses. In fact, the term "propensity score" does not appear. The paper that introduced the propensity score, Rosenbaum and Rubin (1983b), and which by now has over 37,000 Google Scholar cites, had only recently been published at that time and had not yet influenced the

econometrics literature.[10] Subsequently, recognizing the importance of modeling both the assignment mechanism and the conditional outcome distribution has spurred the development of various doubly robust methods. First introduced by Scharfstein et al. (1999), these methods are now generally viewed as the most attractive methods based on unconfoundedness in practice. The incorporation of machine learning techniques into causal inference has further enabled this by reducing the need for *ad hoc* specification searches.

Fourth, effective methods have emerged for other estimands, such as conditional average treatment effects for the treated (CATT)—conditional on (possibly high-dimensional) covariates—and quantile treatment effects. In many cases, decision-makers seek to understand not just the average effects for the entire population but also the extent and nature of effect heterogeneity, or even to estimate personalized assignment rules. Do some subpopulations benefit more from the treatment than others? Do some experience negative effects? The availability of large datasets has led to the development of effective ways of estimating heterogeneous treatment effects (*e.g.*, Wager and Athey 2018).

Finally, scholars have come to realize the importance of establishing the credibility of estimates through validation exercises, particularly placebo analyses. LaLonde (1986) did some placebo analyses looking at the estimated effect for lagged earnings, but primarily focused on the comparisons between nonexperimental and experimental estimates of average treatment effects. The recent empirical literature has placed significant emphasis on supporting main estimates with supplementary analyses, which often take the form of placebo analyses that present estimates of causal effects known to be zero.

Based on these lessons, we offer the following concrete recommendations to practitioners:

- Begin analyses of causal effects with an effort to understand the assignment mechanism. A clear grasp of the "design" is crucial for the credibility of the unconfoundedness assumption.
- Estimate the propensity score using a flexible method. Assess overlap by plotting the distributions of propensity scores for treated and control units. Trim the data based on the propensity score to make the groups more comparable.
- Apply modern methods, such as doubly-robust estimators, to estimate the average causal effects. Explore alternative estimands, such as the conditional average treatment effects and quantile treatment effects.
- Perform placebo tests, such as those using pretreatment outcomes, to validate uncon-

[10]The first mention of the propensity score in the econometrics literature appears to be Card and Sullivan (1988). Interestingly, they cite Rosenbaum and Rubin (1984) rather than the original propensity score paper Rosenbaum and Rubin (1983b).

foundedness. Conduct sensitivity analyses to gauge the robustness of the findings.

We also provide a detailed online tutorial with `R` code to assist researchers in implementing these methods.

References

Alberto Abadie and Matias D Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503, 2018.

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Alberto Abadie and Guido W Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008.

Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011.

Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2008.

Joshua D Angrist and Jorn-Steffen Pischke. 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, 2010.

Dmitry Arkhangelsky and Guido Imbens. Causal models for longitudinal and panel data: A survey. *The Econometrics Journal*, 2024.

Orley Ashenfelter. Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, pages 47–57, 1978.

Orley Ashenfelter and David Card. Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4):648–660, 1985.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey, Guido W Imbens, and Stefan Wager. 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, 2018.

Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Susan Athey, Guido W Imbens, Jonas Metzger, and Evan Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 2021.

Burt S Barnow, Glen George Cain, Arthur Stanley Goldberger, et al. *Issues in the analysis of selectivity bias*. University of Wisconsin, Inst. for Research on Poverty, 1980.

Marianne P Bitler, Jonah B Gelbach, and Hilary W Hoynes. What mean impacts miss: Distributional effects of welfare reform experiments. *The American Economic Review*, 96(4):988–1012, 2006.

Sebastian Calónico and Jeffrey Smith. The women of the national supported work demonstration. *Journal of Labor Economics*, 35(S1):S65–S97, 2017.

David Card and Daniel Sullivan. Measuring the effect of subsidized training programs on movements in and out of employment. *Econometrica*, 1988.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review, Papers and Proceedings*, 107(5):261–65, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), 2018.

Victor Chernozhukov, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. Applied causal inference powered by ml and ai. *arXiv preprint arXiv:2403.02467*, 2024.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, page 00491241221099552, 2022.

Richard K Crump, V Joseph Hotz, Guido Imbens, and Oscar Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, 2006.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, pages 187–199, 2009.

Scott Cunningham. *Causal inference: The mixtape*. Yale University Press, 2018.

Janet Currie, Henrik Kleven, and Esmée Zwiers. Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings*, volume 110, pages 42–48, 2020.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448): 1053–1062, 1999.

Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

Peng Ding. *A first course in causal inference*. CRC Press, 2024.

Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75 (1):259–276, 2007.

Thomas Fraker and Rebecca Maynard. The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, pages 194–227, 1987.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

James J. Heckman. Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46(4):931–959, 1978.

James J Heckman and V Joseph Hotz. 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):862–874, 1989.

James J Heckman, V Joseph Hotz, and Marcelo Dabos. Do we need experimental data to evaluate the impact of manpower training on earnings? *Evaluation Review*, 11(4):395–427, 1987.

James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654, 1997.

James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation

estimator. *The review of economic studies*, 65(2):261–294, 1998.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81 (396):945–960, 1986.

Martin Huber. *Causal analysis: Impact evaluation and Causal Machine Learning with applications in R*. MIT Press, 2023.

Nick Huntington-Klein. *The effect: An introduction to research design and causality*. CRC Press, 2021.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, pages 1–29, 2004.

Guido W Imbens. Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2):373–419, 2015.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

Guido W Imbens, Donald B Rubin, and Bruce I Sacerdote. Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review*, pages 778–794, 2001.

Patrick Kline. Oaxaca-blinder as a reweighting estimator. *American Economic Review*, 101(3): 532–537, 2011.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

Edward E Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1): 31–43, 1983.

Michael Lechner. Earnings and employment effects of continuous gff-the-job training in east germany after unification. *Journal of Business & Economic Statistics*, 17(1):74–90, 1999.

Michael Lechner. Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics*, 84(2):205–220, 2002.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Jerzey Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1923/1990.

James M Robins and Yaacov Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16, 1997.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Paul R Rosenbaum. 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):656–666, 1984.

Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 212–218, 1983a.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983b.

Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387): 516–524, 1984.

Paul R Rosenbaum et al. The role of a second control group in an observational study. *Statistical Science*, 2(3):292–306, 1987.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Donald B Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

DO Scharfstein, A Rotnitzky, and JM Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. comments and rejoinder. *Journal of the American Statistical Association*, 94(448):1121–1146, 1999.

Jeffrey A Smith and Petra E Todd. Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review*, 91(2):112–118, 2001.

Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353, 2005.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Stefan Wager. *Causal Inference: A Statistical Learning Approach*. Cambridge University Press, 2024.

Stefan Wager and Susan Athey. Causal random forests. *arXiv preprint*, 2015.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Yiqing Xu. Causal inference with time-series cross-sectional data: a reflection. *Available at SSRN 3979613*, 2023.

Qingyuan Zhao and Daniel Percival. Entropy balancing is doubly robust, 2016.

Jose R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015. doi: 10.1080/01621459.2015.1023805.

José R Zubizarreta, Elizabeth A Stuart, Dylan S Small, and Paul R Rosenbaum. *Handbook of Matching and Weighting Adjustments for Causal Inference*. CRC Press, 2023.

Appendix A. Tables 4-6 in LaLonde (1986)

The following tables are adapted from Tables 4, 5, and 6 in LaLonde (1986). We thank Robert LaLonde's estate for allowing us to include these tables in the Appendix.

TABLE 4—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW AFDC PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–79 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Adjusted[c] (3)	NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1979 Unadjusted (4)	NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1979 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–79 Treatments Less Comparisons: Without Age (6)	Difference in Differences: Difference in Earnings Growth 1975–79 Treatments Less Comparisons: With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–79: Unadjusted (8)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–79: Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings: Without AFDC (10)	Controlling for All Observed Variables and Pre-Training Earnings: With AFDC (11)
	2,942	−17	−22	851	861	833	883	843	864	854	–
Controls	(220)	(122)	(122)	(307)	(306)	(323)	(323)	(308)	(306)	(312)	
PSID-1	713	−6,443	−4,882	−3,357	−2,143	3,097	2,657	1746	1,354	1664	2,097
	(210)	(326)	(336)	(403)	(425)	(317)	(333)	(357)	(380)	(409)	(491)
PSID-2	1,242	−1,467	−1,515	1,090	870	2,568	2,392	1,764	1,535	1,826	–
	(314)	(216)	(224)	(468)	(484)	(473)	(481)	(472)	(487)	(537)	
PSID-3	665	−77	−100	3,057	2,915	3,145	3,020	3,070	2,930	2,919	–
	(351)	(202)	(208)	(532)	(543)	(557)	(563)	(531)	(543)	(592)	
PSID-4	928	−5,694	−4,976	−2,822	−2,268	2,883	2,655	1,184	950	1,406	2,146
	(311)	(306)	(323)	(460)	(491)	(417)	(434)	(483)	(503)	(542)	(652)
CPS-SSA-1	233	−6,928	−5,813	−3,363	−2,650	3,578	3,501	1,214	1,127	536	1,041
	(64)	(272)	(309)	(320)	(365)	(280)	(282)	(272)	(309)	(349)	(503)
CPS-SSA-2	1,595	−2,888	−2,332	−683	−240	2,215	2,068	447	620	665	–
	(360)	(204)	(256)	(428)	(536)	(438)	(446)	(468)	(554)	(651)	
CPS-SSA-3	1,207	−3,715	−3,150	−1,122	−812	2,603	2,615	814	784	−99	1,246
	(166)	(226)	(325)	(311)	(452)	(307)	(328)	(305)	(429)	(481)	(720)
CPS-SSA-4	1,684	−1,189	−780	926	756	2,126	1,833	1,222	952	827	–
	(524)	(249)	(283)	(630)	(716)	(654)	(663)	(637)	(717)	(814)	

[a] The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1979. Based on the experimental data, an unbiased estimate of the impact of training presented in col. 4 is $851. The first three columns present the difference between each comparison group's 1975 and 1979 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d] See Table 2 for definitions of the comparison groups.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	Pre-Training Year, 1975 Adjusted[c] (3)	Post-Training Year, 1978 Unadjusted (4)	Post-Training Year, 1978 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons Without Age (6)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78 Unadjusted (8)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78 Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings (10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

[a] The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is $886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d] See Table 3 for definitions of the comparison groups.

TABLE 6—ESTIMATED TRAINING EFFECTS USING TWO-STAGE ESTIMATOR

Variables Excluded from the Earnings Equation, but Included in the Participation Equation	Comparison Group	NSW AFDC Females		NSW Males	
		Heckman Correction for Program Participation Bias, Using Estimate of Conditional Expectation of Earnings Error as Regressor in Earnings Equation			
		Estimate of Coefficient for			
		Training Dummy	Estimate of Expectation	Training Dummy	Estimate of Expectation
Marital Status, Residency in an SMSA, Employment Status in 1976, AFDC Status in 1975, Number of Children	*PSID*-1	1,129	−894	−1,333	−2,357
		(385)	(396)	(820)	(781)
	CPS-SSA-1	1,102	−606	−22	−1,437
		(323)	(480)	(584)	(449)
	NSW Controls	837	−18	899	−835
		(317)	(2376)	(840)	(2601)
Employment Status in 1976, AFDC Status in 1975, Number of Children	*PSID*-1	1,256	−823	–	–
		(405)	(410)		
	CPS-SSA-1	439	−979	–	–
		(333)	(481)		
	NSW Controls	–	–	–	–
Employment Status in 1976, Number of Children	*PSID*-1	1,564	−552	−1,161	−2,655
		(604)	(569)	(864)	(799)
	CPS-SSA-1	552	−902	13	−1,484
		(514)	(551)	(584)	(450)
	NSW Controls	851	147	889	−808
		(318)	(2385)	(841)	(2603)
No Exclusion Restrictions	*PSID*-1	1,747	−526	−667	−2,446
		(620)	(568)	(905)	(806)
	CPS-SSA-1	805	−908	213	−1,364
		(523)	(548)	(588)	(452)
	NSW Controls	861	284	889	−876
		(318)	(2385)	(840)	(2601)

Notes: The estimated training effects are in 1982 dollars. For the females, the experimental estimate of impact of the supported work program was $851 with a standard error of $317. The one-step estimates from col. 11 of Table 4 were $2,097 with a standard error of $491 using the *PSID*-1 as a comparison group, $1,041 with a standard error of $503 using the *CPS-SSA*-1 as a comparison group, and $854 with a standard error of $312 using the NSW controls as a comparison group. Estimates are missing for the case of three exclusions using the NSW controls since AFDC status in 1975 cannot be used as an instrument for the NSW females. For the males, the experimental estimate of impact of the supported work program was $886 with a standard error of $476. The one-step estimates from col. 10 of Table 5 were $−1,228 with a standard error of $896 using the *PSID*-1 as a comparison group, $−805 with a standard error of $484 using the *CPS-SSA*-1 as a comparison group, and $662 with a standard error of $506 using the NSW controls as a comparison group. Estimates are missing for the case of three exclusions for the NSW males as AFDC status is not used as an instrument in the analysis of the male trainees.