Online Supplementary Materials *for* Imbens and Xu (2025)

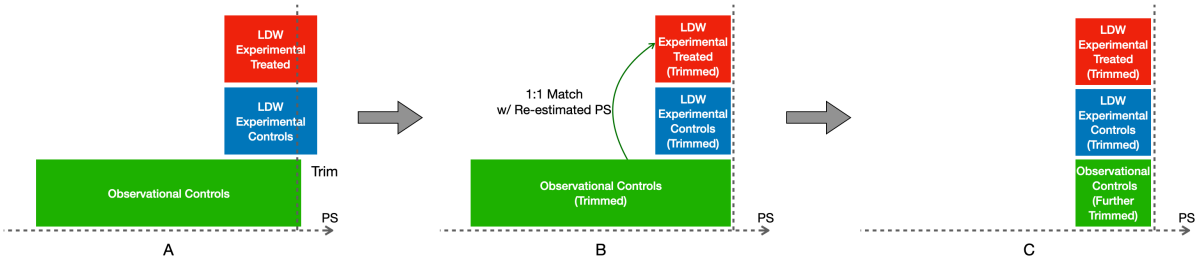
Comparing Experimental and Nonexperimental Methods: What Lessons Have We Learned Four Decades After LaLonde (1986)?

Table of Contents

A.1. Additional Results Using the LaLonde-Dehejia-Wahba (LDW) Data

Trimming to improve overlap. Based on LDW-CPS and LDW-PSID, we further construct two trimmed samples to improve overlap. Trimming involves two steps. First, we merge the experimental controls from LDW-Experimental into LDW-CPS (or LDW-PSID) and estimate each unit's propensity of being included in the experiment using GRF.[A1] We

FIGURE A1. Trimming the LDW Data



Note: A demonstration of the trimming procedure.

set a threshold to trim data with estimated propensity scores exceeding this value. The thresholds for LDW-CPS and LDW-PSID, chosen based on the numbers of available controls units, are 0.9 and 0.8, respectively.[A2] The transition from Step (A) to (B) in Figure A1 illustrates this step. We conduct this procedure to trim all three samples simultaneously to improve overlap in the final samples. This is because merely trimming the nonexperimental controls is inadequate, as the nonexperimental datasets lack particular profiles of participants of the experiment. For LDW-CPS, 22 treated units (12%) are excluded, whereas for LDW-PSID, 105 treated units (57%) are dropped. This underscores the large differences between the nonexperimental controls and the experiment participants based on covariate distributions.

Second, using the trimmed data and the same set of covariates, we re-estimate propensity scores using GRF, this time excluding the experimental controls. We then employ a 1:1 matching based on the re-estimated propensity scores to further trim the nonexperimental controls. This step is illustrated by the progression from Step (B) to (C) in Figure A1. This procedure yields two samples: trimmed LDW-CPS (or trimmed LDW-PSID), composed of the trimmed experimental treated units and trimmed nonexperimental controls, and another trimmed experimental sample, consisting of trimmed experimental treated units and controls. The latter serves as an experimental benchmark for the former. As shown in Figure 1 in the main text, overlap improves significantly in both samples post-trimming, though this comes with the cost of reduced

[A1] In other words, both experimental treated units and controls are labeled as 1 and the nonexperimental units are labeled as 0. This approach is clearly not feasible when experimental controls are unavailable; our objective is to establish experimental benchmarks for the trimmed samples. We include the full set of covariates, including real earnings and unemployment status for the years 1974 and 1975.

[A2] Using CPS-SSA-1 as controls, only four control units have $\hat{e} > 0.9$; using PSID-1 as controls, only nine control units have $\hat{e} > 0.8$.

sample sizes. We conduct similar procedures to the LaLonde male samples and the reconstructed female samples.

ATT and placebo estimates. Table A1 shows the ATT estimates using four different samples: LDW-CPS, LDW-PSID, trimmed LDW-CPS, and trimmed LDW-PSID. The ATT estimates based on the experimental benchmarks are highlighted in bold font in the first row. These estimates are visualized in Figure 2 in the main text.

TABLE A1. ATT Estimates under Unconfoundedness: LDW Samples

	LDW-CPS		LDW-PSID		LDW-CPS (PS Trimmed)		LDW-PSID (PS Trimmed)	
	(1)		(2)		(3)		(4)	
Experimental Benchmark	**1794**	(671)	**1794**	(671)	**1911**	(738)	**306**	(986)
Difference-in-Means	-8498	(582)	-15205	(656)	1484	(824)	-1505	(1220)
Regression	1066	(627)	4	(854)	1751	(824)	-1940	(1154)
Regression w/ Interactions	1133	(624)	688	(635)	1507	(672)	-1511	(900)
GRF	1074	(630)	837	(635)	1460	(665)	-1429	(792)
Nearest Neighbor Matching	1729	(815)	2255	(1404)	2138	(946)	-1565	(1210)
IPW	1224	(690)	665	(899)	1398	(820)	-2038	(1263)
CBPS	1410	(655)	2438	(878)	1471	(813)	-1885	(1370)
Entropy Balancing	1406	(655)	2420	(877)	1472	(813)	-1777	(1552)
DML-ElasticNet	1023	(627)	45	(797)	1803	(803)	-1806	(1050)
AIPW-GRF	1550	(721)	1512	(780)	1440	(820)	-1957	(1158)

Note: ATT estimates using the LDW data. The outcome variable is `re78`. We adjust for the following covariates: `age`, `education`, `black`, `hispanic`, `married`, `nodegree`, `re74`, `u74`, `re75`, and `u75`. The trimmed samples are based on 1:1 matching on propensity scores estimated via GRF. Robust standard errors are in the parentheses.

Table A2 shows the results of the placebo analyses using earnings in 1975 (`re75`) as the placebo outcome and remove both `re75` and `u75` from the set of conditioning variables. These estimates are visualized in Figure 4 in the main text.

TABLE A2. Placebo Test: '75 Earnings as the Outcome

	LDW-CPS		LDW-PSID		LDW-CPS (PS Trimmed)		LDW-PSID (PS Trimmed)	
	(1)		(2)		(3)		(4)	
Experimental Benchmark	**265**	(305)	**265**	(305)	**274**	(334)	**-210**	(693)
Difference-in-Means	-12119	(247)	-17531	(361)	-1457	(485)	-4671	(1057)
Regression	-1135	(272)	-2757	(589)	-1257	(330)	-3695	(853)
Regression w/ Interactions	-1097	(395)	-2641	(367)	-1232	(402)	-3529	(828)
GRF	-1587	(373)	-4347	(343)	-1358	(361)	-3869	(658)
Nearest Neighbor Matching	-1466	(352)	-1914	(805)	-1411	(357)	-3790	(930)
IPW	-1562	(336)	-3285	(736)	-1689	(500)	-4229	(1019)
CBPS	-1229	(298)	-2285	(834)	-1231	(468)	-3676	(1060)
Entropy Balancing	-1228	(298)	-2251	(842)	-1231	(469)	-3552	(1063)
DML-ElasticNet	-1106	(262)	-2484	(499)	-1264	(341)	-3456	(861)
AIPW-GRF	-1265	(263)	-2345	(617)	-1415	(377)	-3675	(810)

Note: ATT estimates using the LDW data. The outcome variable is `re75`. We adjust for the following covariates: `age`, `education`, `black`, `hispanic`, `married`, `nodegree`, `re74`, and `u74`. The trimmed samples are based on 1:1 matching on propensity score estimated via GRF. Robust standard errors are in the parentheses.

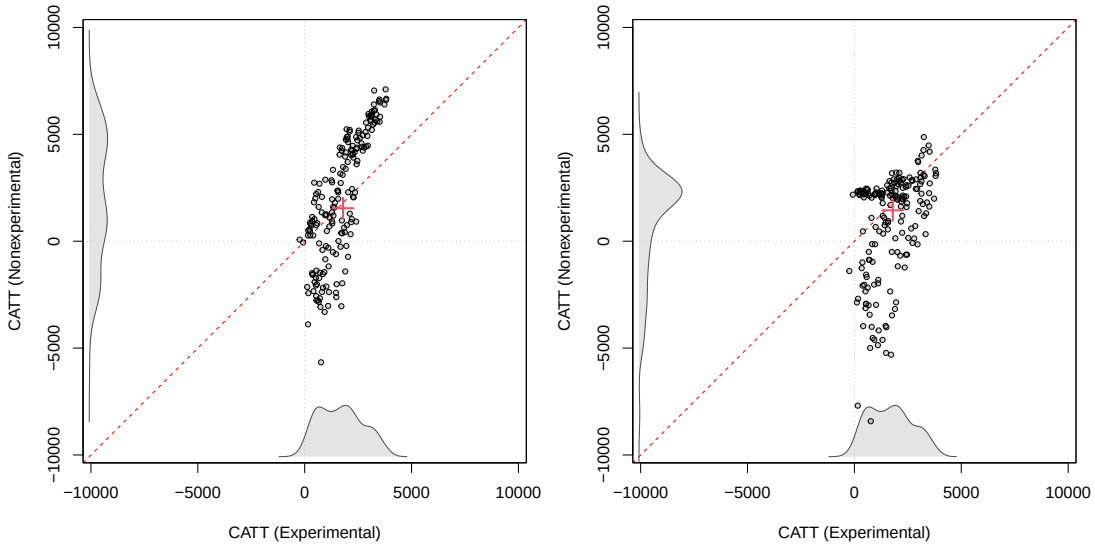
CATT and quantile treatment effects. Exploring causal estimates for alternative estimands, such as heterogeneous treatment effects and quantile treatment effects, can shed light on the plausibility of unconfoundedness. Using both the original LDW data and the trimmed versions, we estimate the CATT using a causal forest through AIPW-GRF. In Figure A2, we plot the estimated CATT from nonexperimental data at the covariate values of each treated unit against their corresponding experimental benchmarks. Each of the four panels uses a different dataset. Every gray dot represents a pair of CATT estimates, while the red cross depicts the pair of estimated ATTs, also estimated via AIPW-GRF. This exercise is, in essence, similar to the test proposed by Athey and Imbens (2015) to explore a finding's robustness to model specifications by splitting samples based on covariate values.

Figure A2 shows that although the AIPW estimator can produce ATT estimates closely aligned with the experimental benchmark using LDW data, its performance for revealing the true CATT is considerably worse. Specifically, with LDW-CPS, CATT estimates span from $-5,693 to $7,364, contrasting with the CATT estimated from experimental data which ranges from $-329 to $3,946. It overestimates CATT that exceed the ATT and underestimates CATT that fall below the ATT. Employing LDW-PSID generates CATT estimates ranging from $-8874 to $4701. With trimmed LDW-CPS, the CATT estimates align more closely with those from the experimental data. However, using trimmed LDW-PSID, the majority of CATT estimates are negative, suggesting significant biases.

We estimate the quantile treatment effects on the treated (QTET) using an IPW approach proposed by Firpo (2007). Figure A3 plots the QTET estimates using both the LDW experimental data and nonexperimental data. The QTET estimates from either the original or trimmed LDW-CPS data align reasonably well with the true QTET, although they are often underpowered. On the other hand, using LDW-PSID data, be it original or trimmed, the estimated QTET display notable biases when compared to the experimental benchmark, which is close to zero.

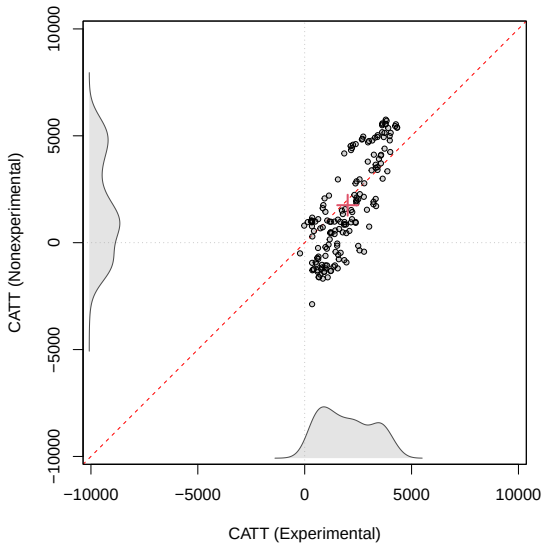
These exercises suggest that, when considering alternative estimands such as CATT and QTET, among the four nonexperimental samples, overall, only trimmed LDW-CPS produces results consistently aligned closely with the experimental benchmarks.

FIGURE A2. CATT Estimates using LDW Data: Experimental vs. Nonexperimental

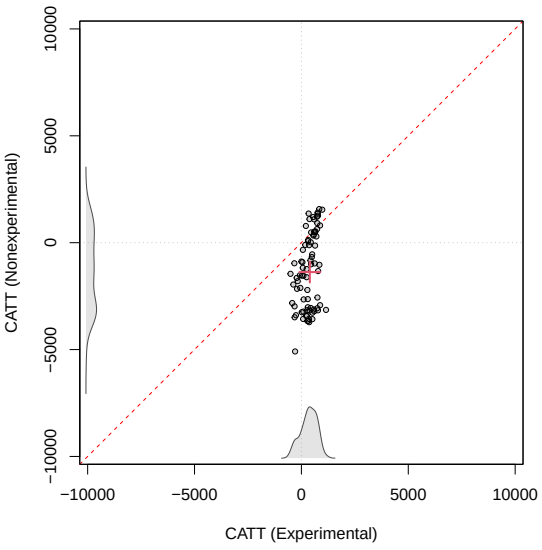


A. LDW-CPS

B. LDW-PSID



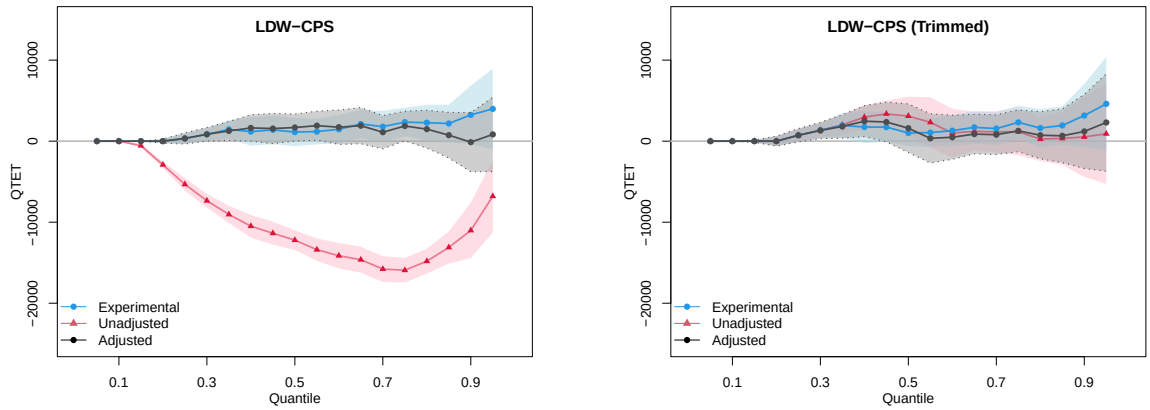
C. LDW-CPS Trimmed



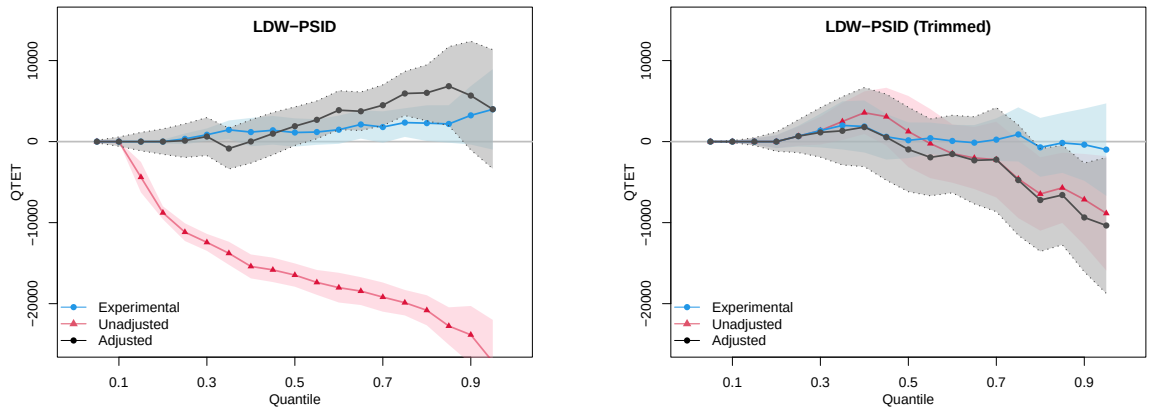
D. LDW-PSID Trimmed

Note: Scatterplots show the CATT using both experimental data (x-axis) and nonexperimental data (y-axis). Each dot corresponds to a CATT estimate based on the covariate values of a treated unit, while each red cross symbolizes the ATT estimates. For every estimate, the AIPW estimator is employed, with the GRF approach for estimating nuisance parameters. Different subfigures indicate various data comparisons: **Subfigure A**: Compares LDW-Experimental with LDW-CPS. **Subfigure B**: Compares LDW-Experimental with LDW-PSID. **Subfigure C**: Compares trimmed LDW-Experimental (removing 22 treated units) against trimmed LDW-CPS. **Subfigure D**: Compares trimmed LDW-Experimental (removing 70 treated units) to trimmed LDW-PSID.

FIGURE A3. Quantile Treatment Effects: Experimental vs. Nonexperimental



A. LDW-CPS

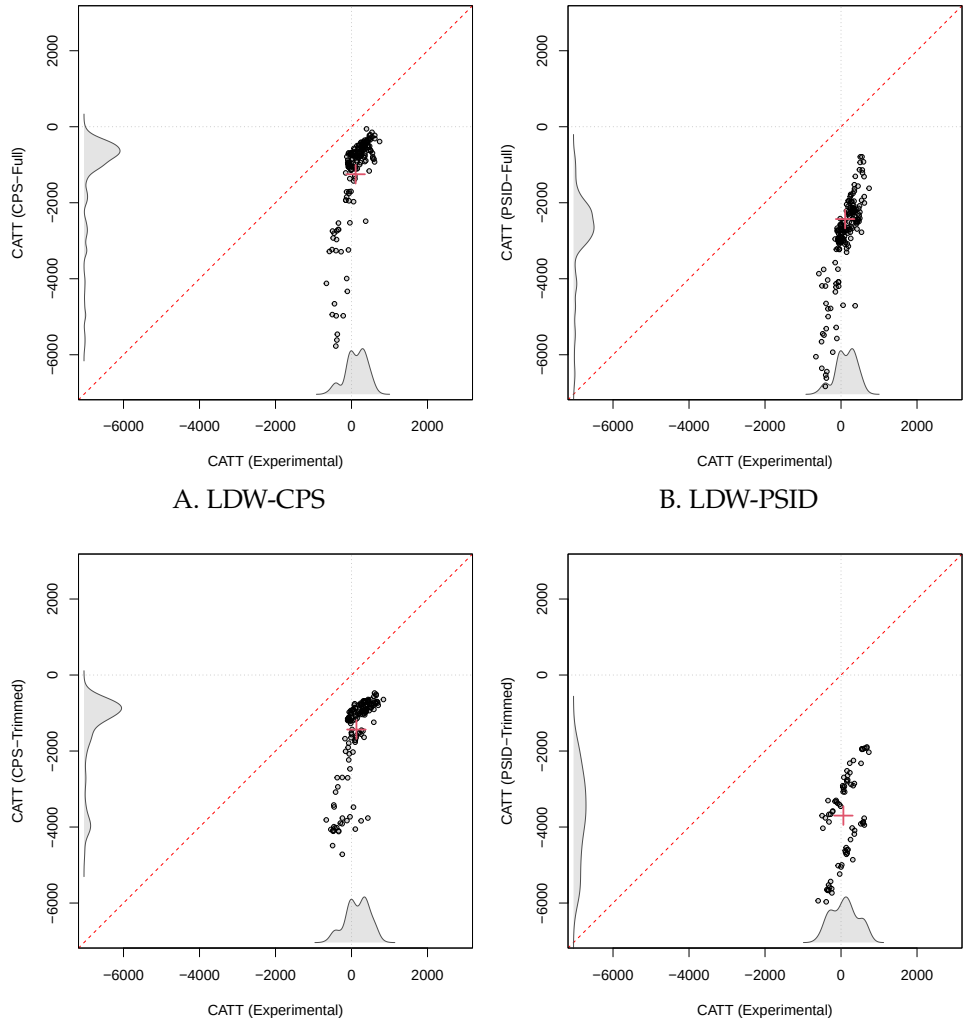


B. LDW-PSID

Note: Figures show the quantile treatment effects on the treated (QTET) using both experimental data (in blue) and nonexperimental data (in red for raw estimates and black for covariate-adjusted estimates). Each dot corresponds to a QTET estimate at a particular quantile, while shaded areas represent bootstrapped 95% confidence intervals. Unadjusted models do not incorporate covariates while adjustment models use the full set of covariates to estimate the propensity scores with a logit. **Subfigure A**: Compares LDW experimental data with LDW-CPS. **Subfigure B**: Compares LDW experimental data with LDW-PSID.

CATT plots for placebo tests. Figure A4 shows that the CATT estimates using 1975 earnings as the placebo outcome.

FIGURE A4. Placebo Tests using LDW Data: Experimental vs. Nonexperimental



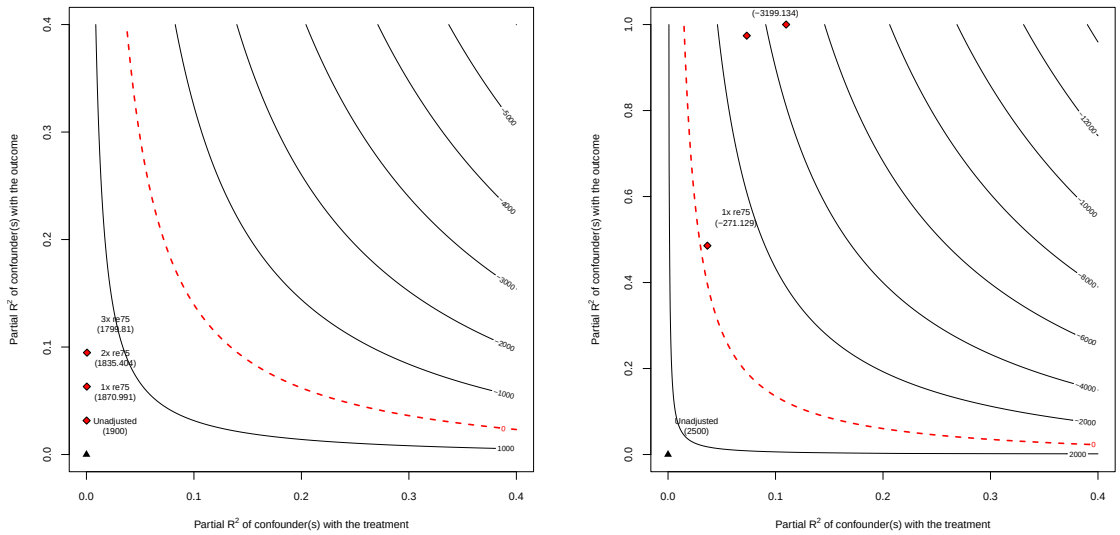
A. LDW-CPS B. LDW-PSID

C. LDW-CPS Trimmed D. LDW-PSID Trimmed

Note: Scatterplots show the CATT on a placebo outcome, real earning in 1975 (`re75`), using both experimental data (x-axis) and nonexperimental data (y-axis). Each dot corresponds to a CATT estimate based on the covariate values of a treated unit, while each red cross symbolizes the ATT estimates. For every estimate, the AIPW estimator is employed, with the GRF approach for estimating nuisance parameters. Different subfigures indicate various data comparisons: **Subfigure A**: Compares LDW-Experimental with LDW-CPS. **Subfigure B**: Compares LDW-Experimental with LDW-PSID. **Subfigure C**: Compares trimmed LDW-Experimental (removing 22 treated units) against trimmed LDW-CPS. **Subfigure D**: Compares trimmed LDW-Experimental (removing 70 treated units) to trimmed LDW-PSID.

Sensitivity analyses. Sensitivity analyses adopt a different approach than placebo tests. Rather than validating unconfoundedness with auxiliary data, they assume unconfoundedness only holds conditional on observed covariates X_i and an unobserved confounder U_i. Rosenbaum and Rubin (1983) suggest modeling the conditional distribution of potential outcomes and the treatment assignment probability (propensity score) given X_i and U_i. While relationships with X_i are estimated, the dependence on U_i is supplied by the researcher to estimate a treatment effect. A causal relationship is considered insensitive to unobserved confounding if the estimated effect remains robust against strong dependence with U_i. Imbens (2003) improves this approach by benchmarking the association between the unobserved U_i and the potential outcomes and treatment with those estimated on observed covariates and introducing contour plots for interpretation. Cinelli and Hazlett (2020) further refine this method by relaxing treatment assignment functional forms and incorporating multiple confounders. Alternatively, in a series of papers (see Rosenbaum (2002) for references), Rosenbaum proposes using the relative odds ratio of propensity scores between treated and control units, examining the range of odds necessary to significantly alter the p-value in a test for a null effect.

FIGURE A5. Sensitivity Analyses for Trimmed LDW-CPS and LDW-PSID



A. LDW-CPS B. LDW-PSID

Note: Contour plot for the treatment effect coefficient $\hat{\tau}_{OLS}$ based on sensitivity analysis first proposed by Imbens (2003) and then modified by Cinelli and Hazlett (2020). The red dashed line indicates $\hat{\tau}_{OLS} = 0$. The benchmark covariate is `re75`. The model is a linear regression with all available long-term covariates included. Both datasets are preprocessed by trimming observations whose estimated propensity scores are smaller than 0.1 or bigger than 0.9.

Figure A5 shows sensitivity analysis results using the trimmed LDW-CPS and LDW-PSID data, with 1975 earnings (`re75`) as the placebo outcome. The estimated training effect from trimmed LDW-CPS is less sensitive to potential confounders than that from trimmed LDW-PSID.

ATT Estimates without using '74 information. Table A3 presents ATT estimates based on the LDW samples without using earnings and employment status in 1974.

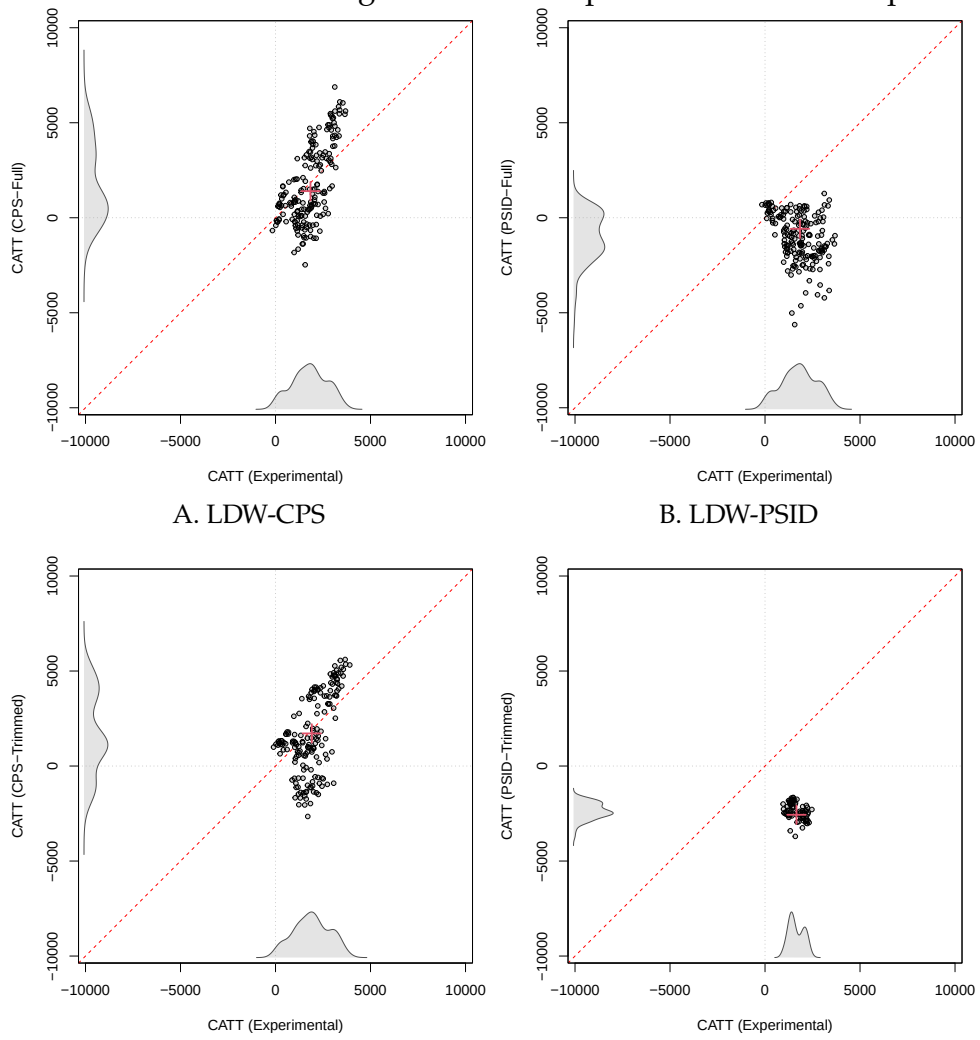
TABLE A3. ATT Estimates without Using `re74` and `u74`: LDW Samples

	LDW-CPS		LDW-PSID		LDW-CPS (PS Trimmed)		LDW-PSID (PS Trimmed)	
	(1)		(2)		(3)		(4)	
Experimental Benchmark	**1794**	(671)	**1794**	(671)	**1917**	(695)	**1480**	(974)
Difference-in-Means	-8498	(582)	-15205	(656)	817	(788)	-3929	(1871)
Regression	1167	(626)	428	(907)	1120	(764)	-3523	(2014)
Regression w/ Interactions	1198	(617)	721	(630)	1169	(630)	-3705	(889)
GRF	703	(617)	-1531	(614)	1194	(621)	-3980	(760)
Nearest Neighbor Matching	1333	(761)	210	(1808)	1353	(857)	-4755	(2078)
IPW	1016	(648)	-570	(1201)	1256	(775)	-4120	(2127)
CBPS	1193	(645)	673	(1136)	1160	(784)	-3735	(2069)
Entropy Balancing	1192	(645)	685	(1142)	1162	(784)	-3728	(2065)
DML-ElasticNet	1122	(622)	404	(812)	959	(761)	-3734	(1903)
AIPW-GRF	1429	(674)	-657	(1079)	1377	(777)	-4062	(2050)

Note: ATT estimates using the LDW data. The outcome variable is `re78`. We adjust for the following covariates: `age`, `education`, `black`, `hispanic`, `married`, `nodegree`, `re75`, and `u75`. The trimmed samples are based on 1:1 matching on propensity scores estimated via GRF. Robust standard errors are in the parentheses.

CATT plots without using '74 information. Figure A6 shows that the CATT estimates without using 1974 earnings and employment status.

FIGURE A6. Placebo Tests using LDW Data: Experimental vs. Nonexperimental



A. LDW-CPS B. LDW-PSID

C. LDW-CPS Trimmed D. LDW-PSID Trimmed

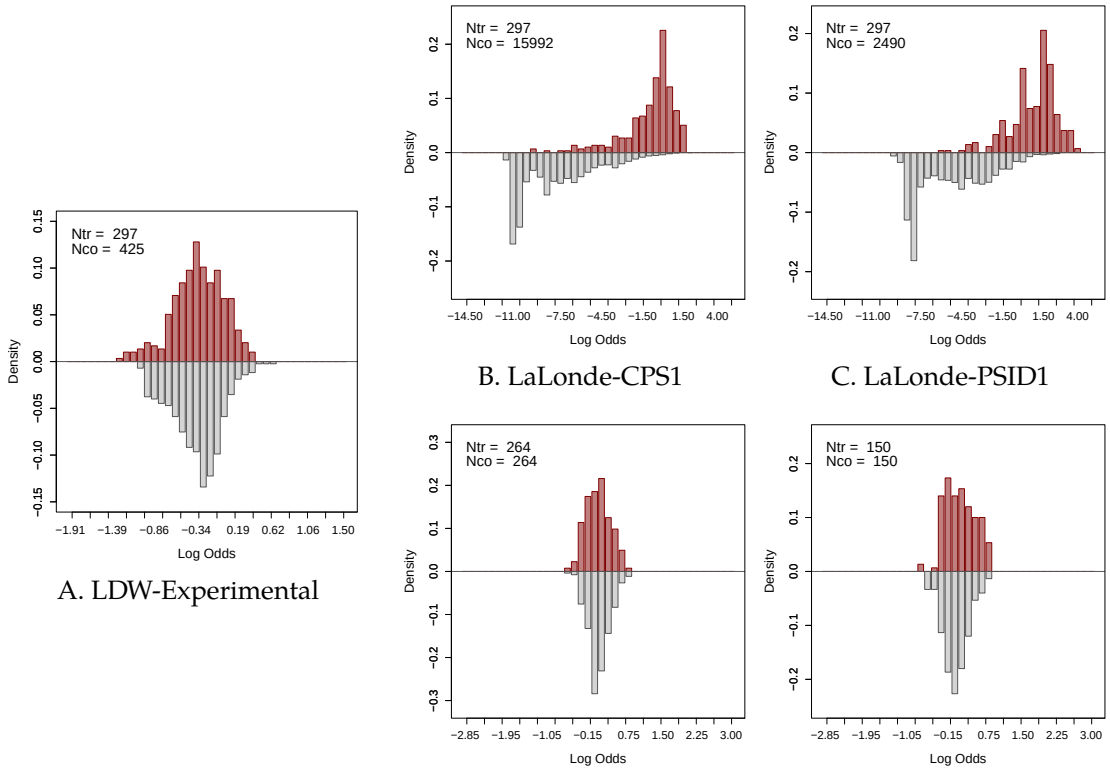
Note: Scatterplots show the CATT on real earning in 1978 (`re78`), using both experimental data (x-axis) and nonexperimental data (y-axis), but without using 1974 earnings and employment status. . Each dot corresponds to a CATT estimate based on the covariate values of a treated unit, while each red cross symbolizes the ATT estimates. For every estimate, the AIPW estimator is employed, with the GRF approach for estimating nuisance parameters. Different subfigures indicate various data comparisons: **Subfigure A**: Compares LDW-Experimental with LDW-CPS. **Subfigure B**: Compares LDW-Experimental with LDW-PSID. **Subfigure C**: Compares trimmed LDW-Experimental (removing 10 treated units) against trimmed LDW-CPS. **Subfigure D**: Compares trimmed LDW-Experimental (removing 106 treated units) to trimmed LDW-PSID.

A.2. Results based on the LaLonde Male Samples

Two pretreatment variables, earnings in 1974 and employment status in 1974 are absent from this sample.

Assessing overlap. Figure A7 demonstrates overlap in the LDW using the propensity score estimated via GRF (log odds ratio).

FIGURE A7. Assessing the Overlap in the LaLonde Data (Male Sample)



A. LDW-Experimental

B. LaLonde-CPS1

C. LaLonde-PSID1

D. Trimmed LaLonde-CPS1

E. Trimmed LaLonde-PSID1

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 number of treated and control units, respectively. **Subfigure A**: Experimental (male sample). **Subfigure B**: LaLonde-CPS1. **Subfigure C**: LaLonde-PSID1. **Subfigure D**: Trimmed LaLonde-CPS1. **Subfigure E**: Trimmed LaLonde-PSID1. For C and D, the propensity scores are reestimated after trimming. Covariates do not include `re74` and `u74`.

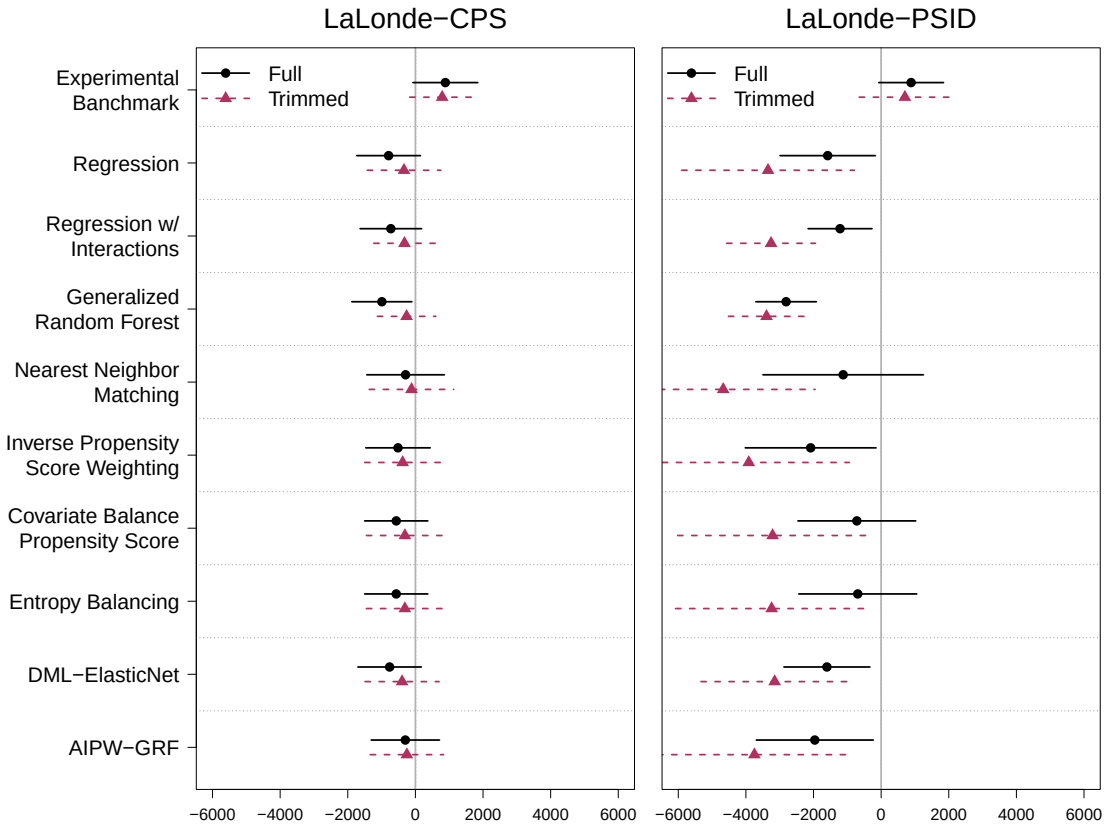
ATT estimates. Table A4 shows the ATT estimates using the original LaLonde male samples, of which the LDW sample is a subset. Two covariates, `re74` and `u74`, are absent from this sample. CPS-SSA-1 and PSID-1 are used as nonexperimental control groups. We also trim the data to improve overlap, using the same procedure applied to the LDW samples. Figure A8 visualizes the ATT estimates.

TABLE A4. ATT Estimates: LaLonde Male Samples

	NSW-CPS		NSW-PSID		NSW-CPS (PS Trimmed)		NSW-PSID (PS Trimmed)	
	(1)		(2)		(3)		(4)	
Experimental Benchmark	**886**	(488)	**886**	(488)	**802**	(486)	**569**	(670)
Difference-in-Means	-8870	(408)	-15578	(508)	-356	(572)	-3468	(1205)
Regression	-792	(480)	-1581	(719)	-369	(563)	-3153	(1331)
Regression w/ Interactions	-726	(463)	-1215	(481)	-521	(444)	-3270	(639)
GRF	-946	(452)	-2782	(455)	-393	(422)	-3255	(566)
Nearest Neighbor Matching	-290	(585)	-1123	(1210)	-330	(651)	-4688	(1394)
IPW	-533	(486)	-2082	(977)	-587	(588)	-3850	(1417)
CBPS	-566	(476)	-719	(888)	-530	(600)	-3258	(1433)
Entropy Balancing	-567	(476)	-692	(889)	-532	(600)	-3337	(1472)
DML-ElasticNet	-762	(478)	-1604	(649)	-399	(563)	-3257	(1151)
AIPW-GRF	-271	(508)	-1880	(873)	-604	(575)	-3897	(1532)

Note: ATT estimates using the orignal LaLonde data (male sample). The control groups are CPS-SSA-1 (CPS1) and PSID-1 (PSID1). The outcome variable is `re78`. We adjust for the following covariates: `age`, `education`, `black`, `hispanic`, `married`, `nodegree`, `re75`, and `u75`. The trimmed samples are based on 1:1 matching on propensity score estimated via GRF. Robust standard errors are in the parentheses.

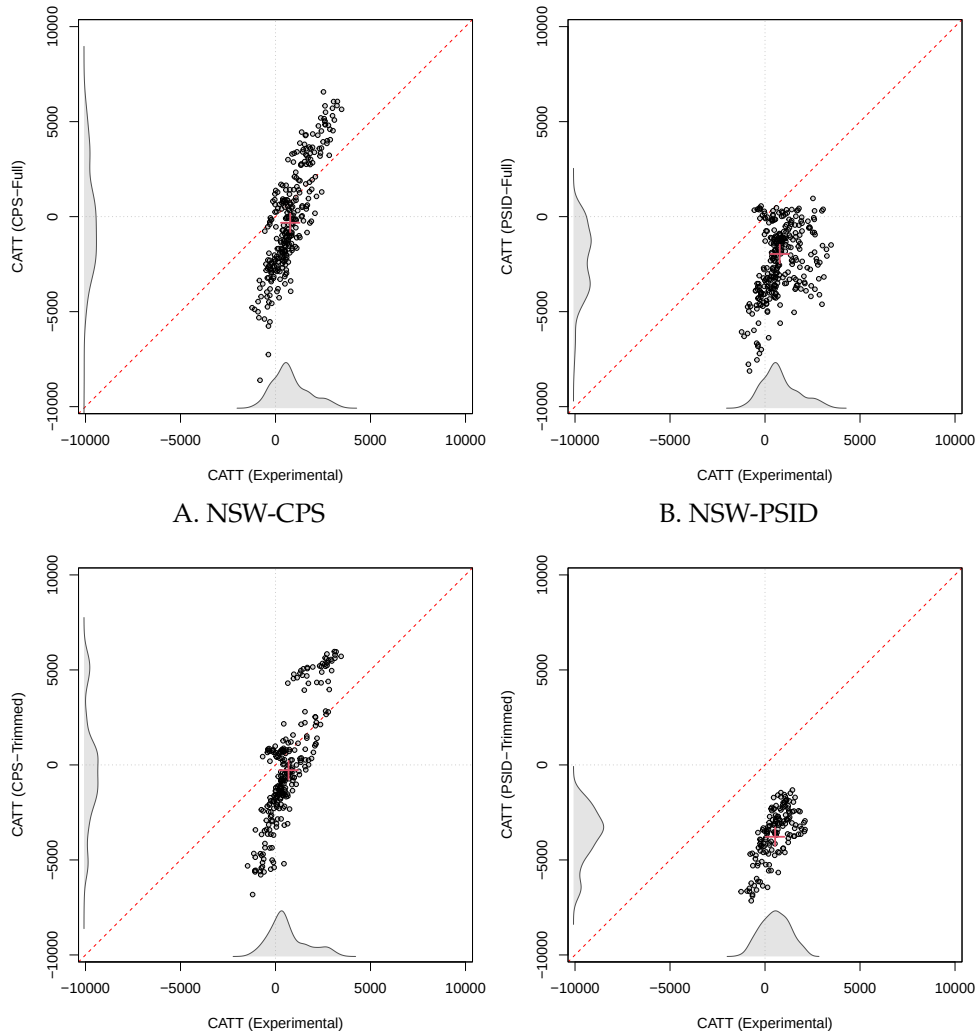
FIGURE A8. ATT Estimates Given Unconfoundedness: LaLonde Male Samples



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

CATT estimates. Figure A9 shows the CATT estimates using the original LaLonde data (male sample). Two covariates, `re74` and `u74`, are not included in this sample.

FIGURE A9. CATT Estimates for the LaLonde Data (Male Sample)



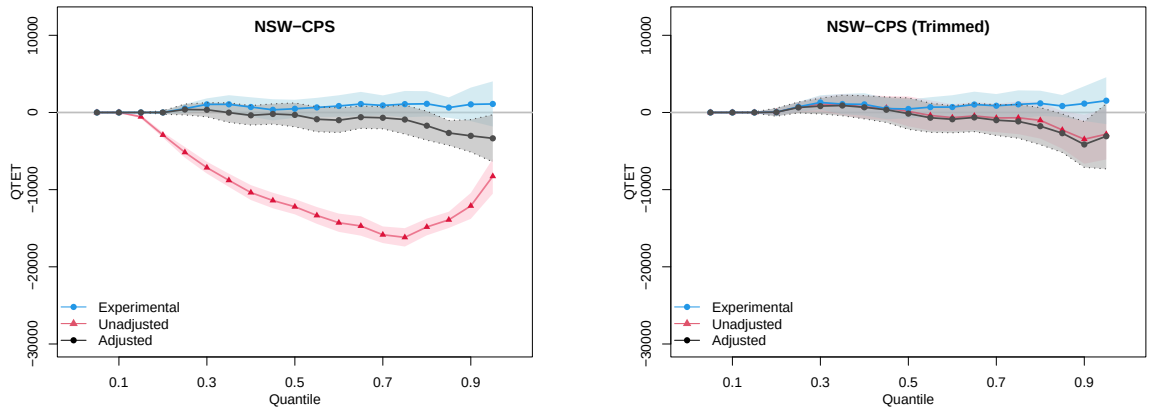
A. NSW-CPS B. NSW-PSID

C. NSW-CPS Trimmed D. NSW-PSID Trimmed

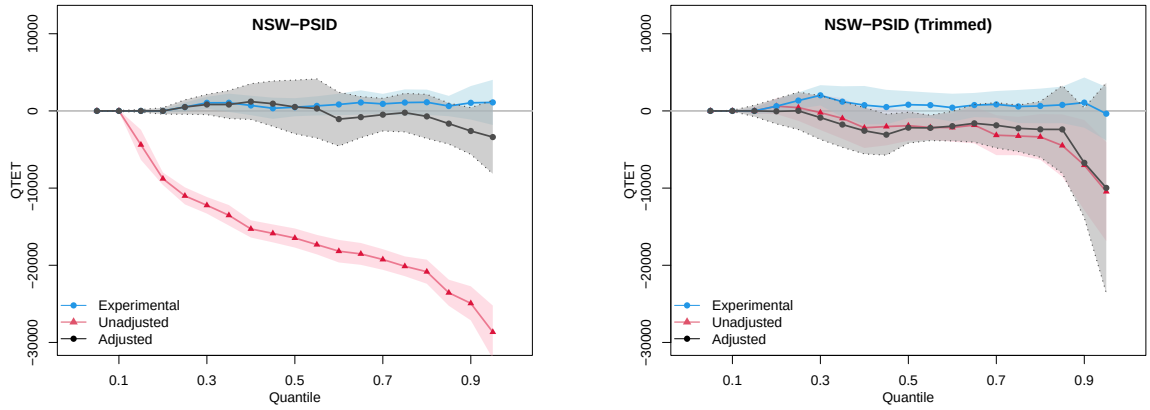
Note: Scatterplots show the CATT using both experimental data (x-axis) and nonexperimental data (y-axis) from LaLonde (1986) (male sample). Each dot corresponds to a CATT estimate based on the covariate values of a treated unit, while each red cross symbolizes the ATT estimates. For every estimate, the AIPW estimator is employed, with the GRF approach for estimating nuisance parameters. Different subfigures indicate various data comparisons: **Subfigure A**: Compares Experimental with LaLonde-CPS1. **Subfigure B**: Compares Experimental with LaLonde-PSID1. **Subfigure C**: Compares trimmed Experimental (removing 30 treated units) against trimmed NSW-CPS. **Subfigure D**: Compares trimmed Experimental (removing 150 treated units) to trimmed NSW-PSID.

Quantile treatment effects. Figures A10 shows the quantile treatment effects on the treated using the original LaLonde (NSW) male samples.

FIGURE A10. Quantile Treatment Effects: Experimental vs. Nonexperimental



A. NSW-CPS

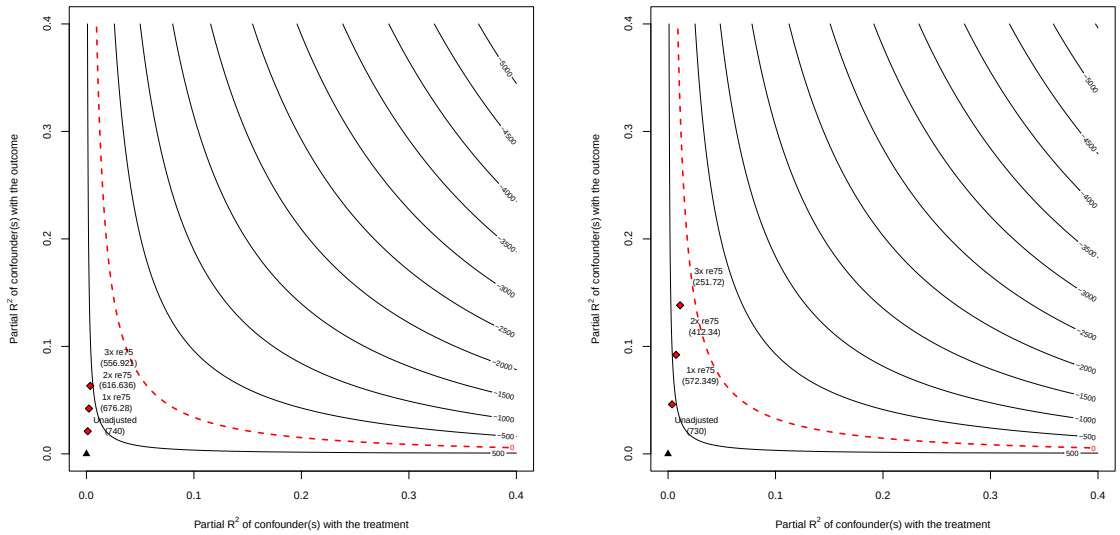


B. NSW-PSID

Note: Figures show the quantile treatment effects on the treated (QTET) using both experimental data (in blue) and nonexperimental data (in red for raw estimates and black for covariate-adjusted estimates). Each dot corresponds to a QTET estimate at a particular quantile, while shaded areas represent bootstrapped 95% confidence intervals. Unadjusted models do not incorporate covariates while adjustment models use the full set of covariates to estimate the propensity scores with a logit. **Subfigure A**: Compares NSW experimental data with NSW-CPS. **Subfigure B**: Compares NSW experimental data with NSW-PSID.

Sensitivity analyses. Figure A11 shows the results of the sensitivity analyses using the trimmed LaLonde male samples, including trimmed NSW-CPS and trimmed NSW-PSID. We used 1975 earnings (`re75`) as the benchmark covariate. The analysis suggests that the estimated training effects are robust to potential confounders that behavior like `re75`. For instance, with trimmed NSW-CPS or NSW-PSID, the estimate remains positive and substantial even when a confounder's correlations with treatment and outcome are triple those of `re75`.

FIGURE A11. Sensitivity Analyses for Trimmed NSW-CPS and NSW-PSID



A. NSW-CPS B. NSW-PSID

Note: Contour plot for the treatment effect coefficient $\hat{\tau}_{OLS}$ based on sensitivity analysis first proposed by Imbens (2003) and then modified by Cinelli and Hazlett (2020). The red dashed line indicates $\hat{\tau}_{OLS} = 0$. The benchmark covariate is `re75`. The model is a linear regression with all available long-term covariates included. Both datasets are preprocessed by trimming observations whose estimated propensity scores are smaller than 0.1 or bigger than 0.9.

A.3. Results based on the Reconstructed LaLonde Female Samples

We report findings using the LaLonde female samples reconstructed by Calónico and Smith (2017), referred to as the LaLonde-Calónico-Smith (LCS) sample. Consistent with LaLonde's original analysis, the outcome variable is earnings in 1979 (`re79`). We use the same set of covariates as in LaLonde (1986). Notably, this set does not include two pretreatment variables: earnings in 1974 and employment status in 1974. We also exclude the number of children in 1975 (`nchildren75`), which is available in the LCS dataset, from the covariates so that it can serve as a placebo outcome. We also trim the data to improve overlap, using the same procedure applied to the LDW samples. The threshold for the propensity score to trim the sample in Step (A) is set at 0.9.

Descriptive statistics. Table A5 shows the descriptive statistics of the reconstructed LaLonde female sample.

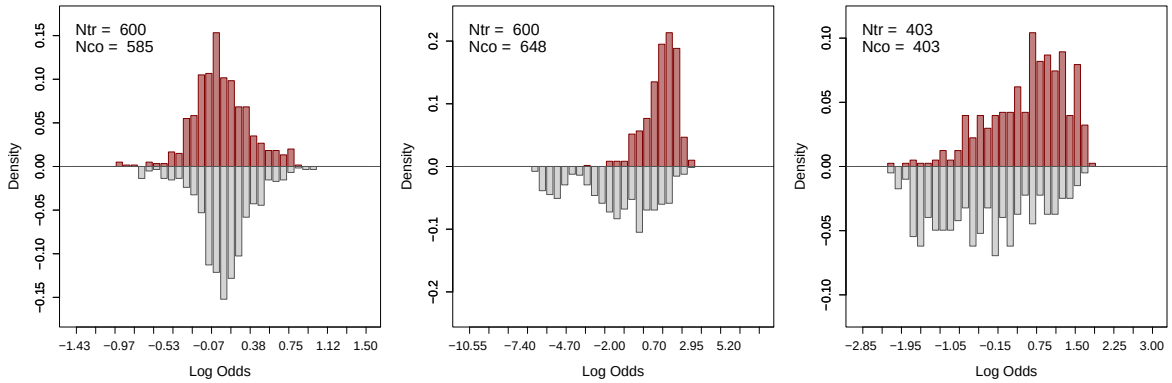
TABLE A5. Descriptive Statistics: LCS Female Samples

	NSW Treated	**NSW Control**	**PSID-1**
	(1)	(2)	(3)
Age	33.77	33.74	37.07
	(7.40)	(7.15)	(10.57)
Years of School	10.31	10.26	11.30
	(1.88)	(2.03)	(2.77)
Proportion High School Dropouts	0.70	0.68	0.45
	(0.46)	(0.47)	(0.50)
Proportion Married	0.02	0.04	0.02
	(0.15)	(0.19)	(0.14)
Proportion Black	0.84	0.82	0.65
	(0.37)	(0.39)	(0.48)
Proportion Hispanic	0.11	0.13	0.02
	(0.32)	(0.33)	(0.12)
Real Eearnings in 1975 (thousand)	0.86	0.88	7.51
	(2.01)	(2.19)	(7.54)
Proportion Unemployed in 1975	0.73	0.75	0.28
	(0.44)	(0.44)	(0.45)
#Observations	**600**	**585**	**648**

Note: Standard deviations are in the parentheses.

Assessing overlap. Figure A12 demonstrates overlap in the LCS using the propensity score estimated via GRF (log odds ratio).

FIGURE A12. Assessing the Overlap in the Reconstructed LaLonde Female Samples



A. LCS-Experimental B. LCS-PSID C. Trimmed LCS-PSID

Note: Histograms depict the log odds ratios, i.e., $\log \frac{\hat{e}}{1-\hat{e}}$, using propensity score estimated through GRF. Each subfigure represents a different sample. **Subfigure A**: LCS-Experimental. **Subfigure B**: LCS-PSID. **Subfigure C**: Trimmed LCS-PSID. For C, the propensity score is reestimated after trimming. Covariates do not include `re74`, `u74`, and `nchildren75`.

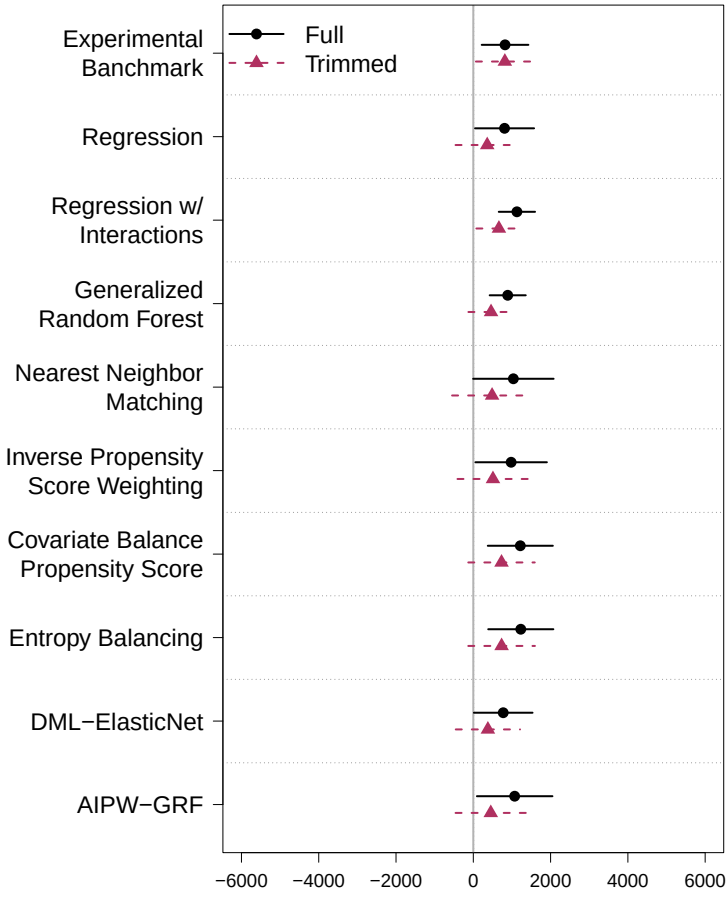
ATT estimates. Table A6 shows the ATT estimates using the reconstructed LaLonde female samples. Reconstructed PSID-1 is used as the nonexperimental control group. Figure A13 visualizes the ATT estimates.

TABLE A6. ATT Estimates: Reconstructed LaLonde Female Samples

	LCS-PSID		LCS-PSID (PS Trimmed)	
	(1)		(2)	
Experimental Benchmark	**821**	(308)	**785**	(380)
Difference-in-Means	-4172	(412)	-804	(422)
Regression	808	(389)	359	(414)
Regression w/ Interactions	1128	(239)	608	(296)
GRF	926	(238)	396	(296)
Nearest Neighbor Matching	1037	(531)	443	(519)
IPW	986	(475)	493	(460)
CBPS	1217	(429)	670	(435)
Entropy Balancing	1229	(430)	673	(436)
DML-ElasticNet	775	(387)	374	(418)
AIPW-GRF	1088	(503)	397	(460)

Note: ATT estimates the reconstructed LaLonde female samples. The control group is PSID-1 (PSID1). The outcome variable is `re79`. The trimmed samples are based on 1:1 matching on propensity score estimated via GRF. Robust standard errors are in the parentheses.

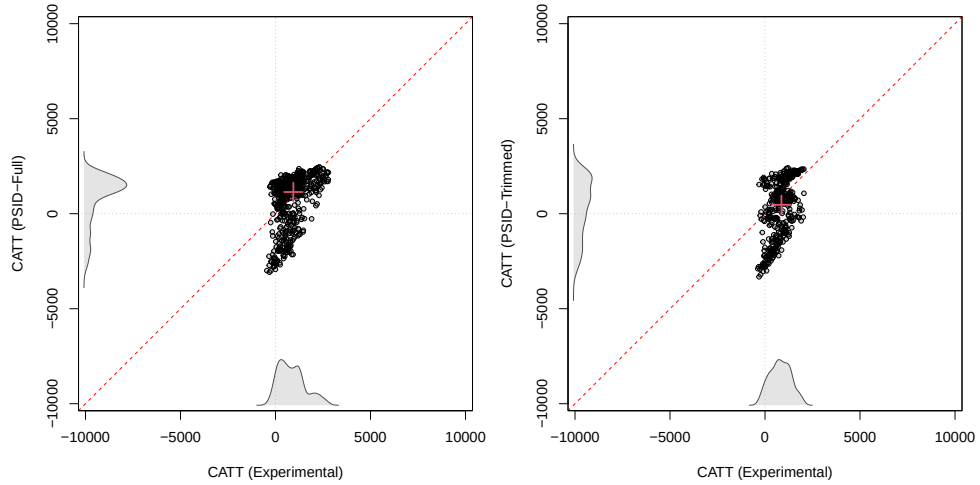
FIGURE A13. ATT Estimates: Reconstructed Female Samples



Note: The above figures show the ATT estimates and their 95% confidence intervals using two different samples: LCS-PSID and Trimmed LCS-PSID. The same ten estimators are employed. The red dashed line and pink band represent the experiment benchmark and its 95% confidence intervals, respectively.

CATT estimates. Figure A14 shows the CATT estimates using the reconstructed LaLonde female samples.

FIGURE A14. CATT Estimates: Reconstructed Female Samples

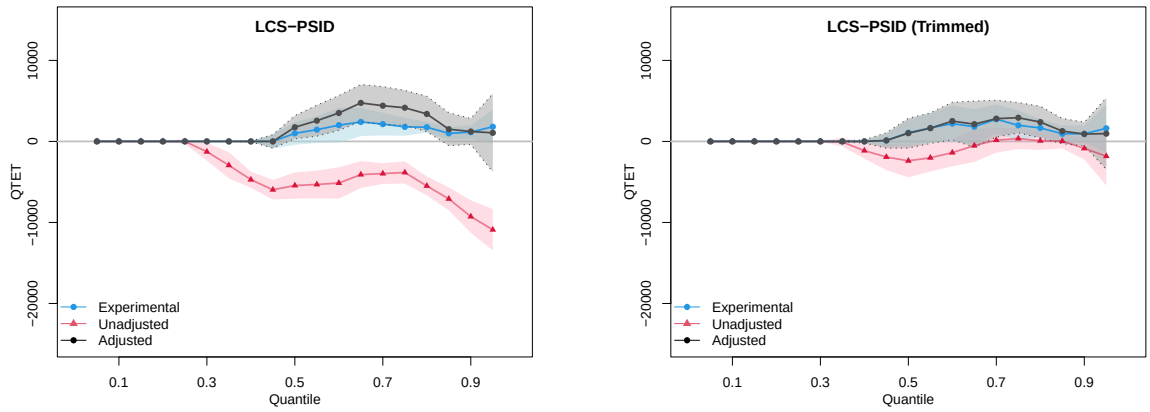


A. LCS-PSID B. LCS-PSID Trimmed

Note: Scatterplots show the CATT using both experimental data (x-axis) and nonexperimental data (y-axis) from the reconstructed LaLonde female samples. Each dot corresponds to a CATT estimate based on the covariate values of a treated unit, while each red cross symbolizes the ATT estimates. For every estimate, the AIPW estimator is employed, with the GRF approach for estimating nuisance parameters. Different subfigures indicate various data comparisons: **Subfigure A**: Compares LCS-Experimental with LaLonde-PSID1. **Subfigure B**: Compares Trimmed LCS-Experimental to Trimmed LCS-PSID.

Quantile treatment effects. Figures A15 shows the quantile treatment effects on the treated using the reconstructed LaLonde female samples.

FIGURE A15. Quantile Treatment Effects: Reconstructed Female Samples



A. NSW-PSID

Note: Figures show the quantile treatment effects on the treated (QTET) using the reconstructed LaLonde female samples. Results from the experimental data are shown in blue and results from the nonexperimental data are shown in red for raw estimates and black for covariate-adjusted estimates. Each dot corresponds to a QTET estimate at a particular quantile, while shaded areas represent bootstrapped 95% confidence intervals. Unadjusted models do not incorporate covariates while adjustment models use the full set of covariates to estimate the propensity scores with a logit.

Placebo Tests. Table A7 shows the results of the placebo analyses using the number of children in 1975 (`nchildren75`) as the placebo outcome.

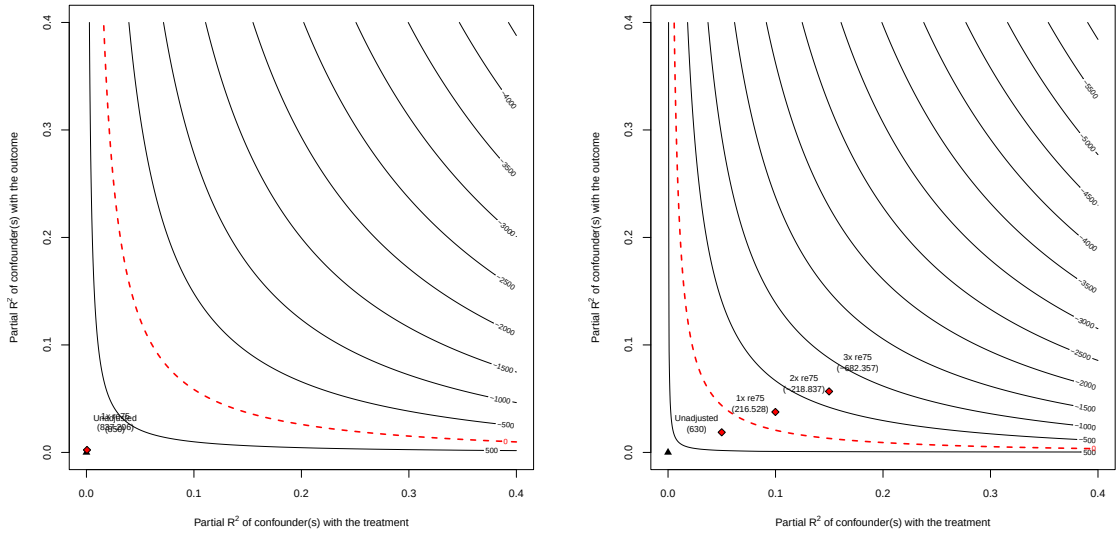
TABLE A7. Placebo Test: Number of Children in 1975 as the Outcome

	LCS-PSID		LCS-PSID (PS Trimmed)	
	(1)		(2)	
Experimental Benchmark	**-0.05**	(0.08)	**-0.08**	(0.09)
Difference-in-Means	0.47	(0.09)	0.14	(0.11)
Regression	-0.18	(0.11)	-0.05	(0.12)
Regression w/ Interactions	-0.22	(0.06)	-0.15	(0.07)
GRF	-0.51	(0.06)	-0.36	(0.07)
Nearest Neighbor Matching	-0.71	(0.14)	-0.49	(0.14)
IPW	-0.68	(0.15)	-0.41	(0.15)
CBPS	-0.27	(0.14)	-0.12	(0.13)
Entropy Balancing	-0.27	(0.14)	-0.12	(0.13)
DML-ElasticNet	-0.13	(0.11)	-0.00	(0.12)
AIPW-GRF	-0.74	(0.15)	-0.49	(0.14)

Note: ATT estimates using the LDW data. The outcome variable is `nchildren75`. The trimmed samples are based on 1:1 matching on propensity score estimated via GRF. Robust standard errors are in the parentheses.

Sensitivity analyses. Figure A16 shows the results of the sensitivity analyses using the trimming LCS data. We used 1975 earnings (`re75`) as the benchmark covariate. The analysis shows that the estimated training effect based on LCS-PSID is sensitive to potential confounders that behave like `re75`.

FIGURE A16. Sensitivity Analyses for Trimmed LCS-CPS and LCS-PSID



A. LCS-Experimental B. LCS-PSID

Note: Contour plot for the treatment effect coefficient $\hat{\tau}_{OLS}$ based on sensitivity analysis first proposed by Imbens (2003) and then modified by Cinelli and Hazlett (2020). The red dashed line indicates $\hat{\tau}_{OLS} = 0$. The benchmark covariate is `re75`. The model is a linear regression with all available long-term covariates included. Both datasets are preprocessed by trimming observations whose estimated propensity scores are smaller than 0.1 or bigger than 0.9.

A.4. Lottery Prizes on Labor Earnings (Imbens-Rubin-Sacerdote Data)

We now turn to the Imbens-Rubin-Sacerdote lottery data (Imbens et al. 2001). 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.

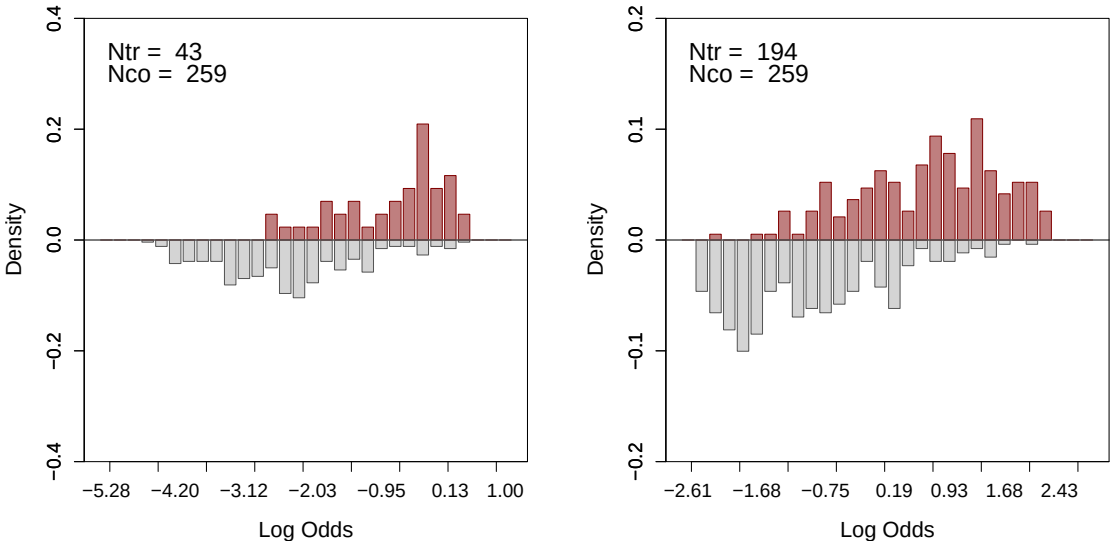
There are three treatment and control groups. The control group, termed "non-winners," consists of 259 season ticket holders who have won a small, one-time prize, ranging from \$100 to \$5,000 (in essence, they are one-time, minor winners). The treatment groups, labeled "big winners" (43 individuals) 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.

Main Findings. 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 A17 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.

FIGURE A17.
Assessing Overlap in the Imbens-Rubin-Sacerdote Lottery Data



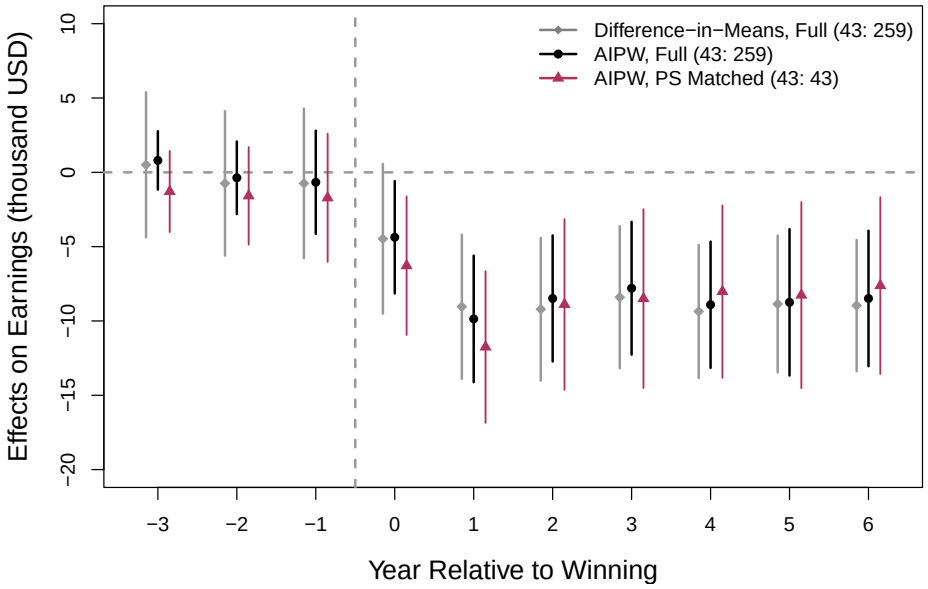
A. Big Winners vs Non-Winners

B. Small Winners vs Non-Winners (Trimmed)

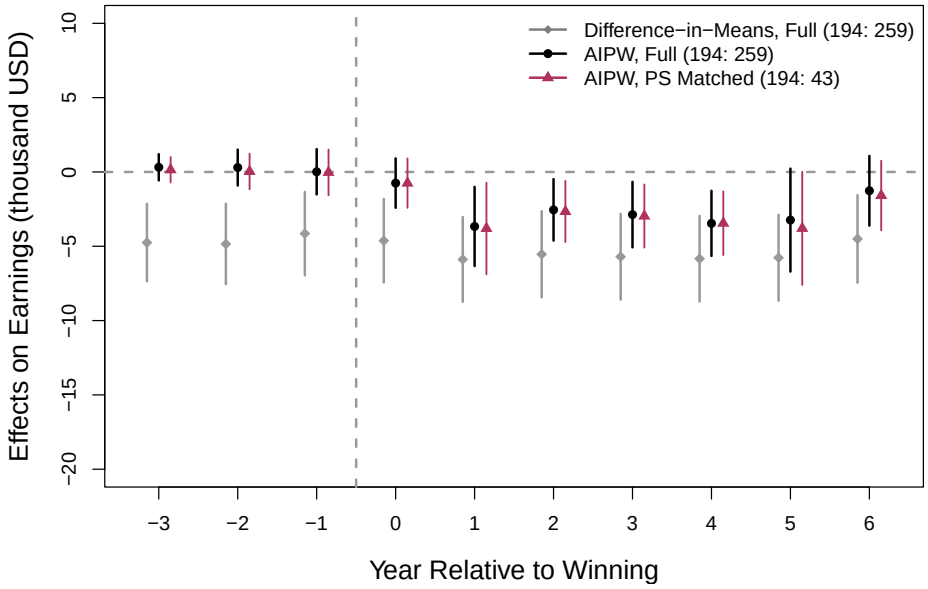
Note: Histograms depict the log odds ratios, i.e., $\log \frac{\hat{e}}{1-\hat{e}}$, using propensity scores estimated through generalized random forest. N_{tr} and N_{co} represent the numbers of treated and control units, respectively.

We estimate the ATT for labor income from Year −3 to Year 6 separately using both difference-in-means and AIPW-GRF. Figure A18 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.

FIGURE A18.
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).

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.

Below we provide additional results using the Imbens-Rubin-Sacerdote lottery data.

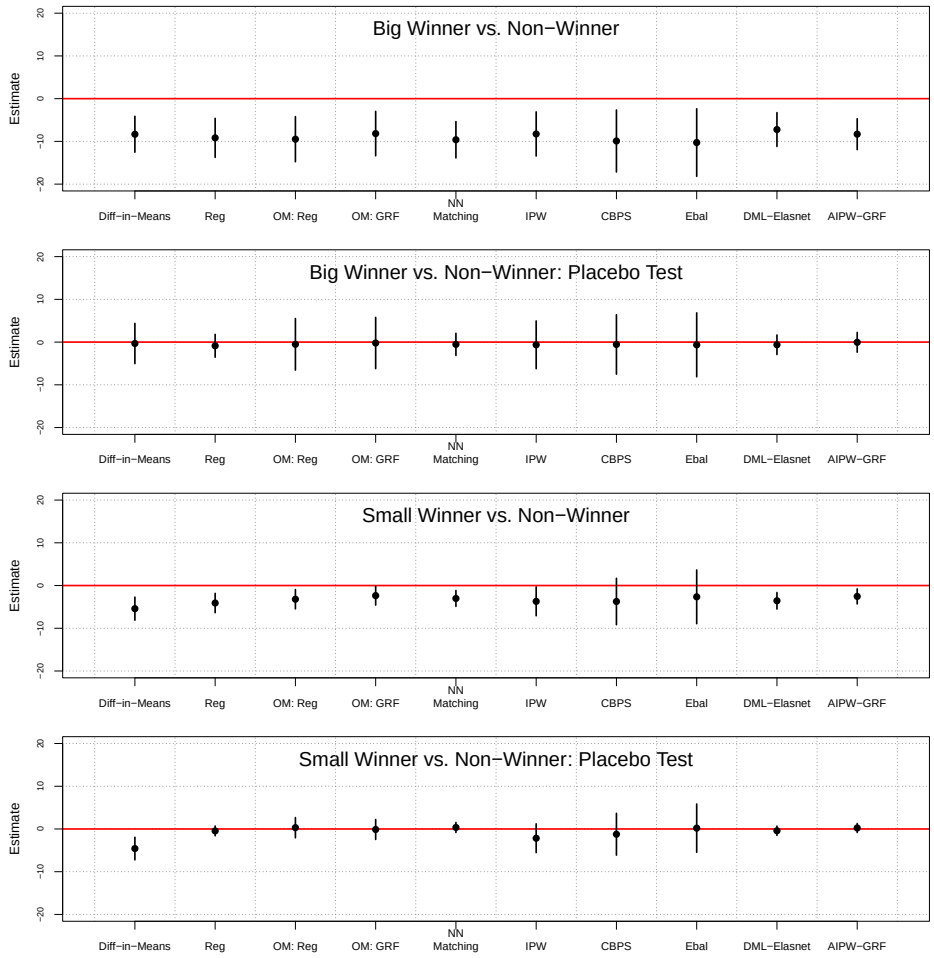
ATT and placebo estimates. Table A8 shows the ATT estimates for the real outcome—average annual labor earnings from Year 0 to Year 6—and the placebo outcome—average annual labor earnings from Year –3 to Year –1—based on the IRS data. These estimates are visualized in Figure A19.

Table A8. ATT and Placebo Estimates: IRS Data

	Winning a Big Prize				Winning a Small Prize			
	Post-Winning Average Earning Years 0:6		Pre-Winning Average Earning Years -3:-1		Post-Winning Average Earning Years 0:6		Pre-Winning Average Earning Years -3:-1	
	(1)		(2)		(3)		(4)	
Difference-in-Means	-8.33	(2.13)	-0.33	(2.39)	-5.41	(1.37)	-4.58	(1.35)
Regression	-9.17	(2.32)	-0.87	(1.36)	-4.09	(1.15)	-0.46	(0.59)
Regression w/ Interactions	-9.49	(2.66)	-0.52	(3.03)	-3.20	(1.15)	0.30	(1.20)
GRF	-8.19	(2.60)	-0.28	(3.02)	-2.40	(1.13)	-0.10	(1.19)
Nearest Neighbor Matching	-9.62	(2.17)	-0.53	(1.32)	-3.02	(0.95)	0.36	(0.60)
IPW	-8.44	(2.68)	-0.79	(2.87)	-3.55	(1.70)	-2.04	(1.71)
CBPS	-9.91	(3.69)	-0.55	(3.53)	-3.74	(2.76)	-1.23	(2.49)
Entropy Balancing	-10.27	(4.02)	-0.64	(3.80)	-2.64	(3.20)	0.20	(2.88)
DML-ElasticNet	-7.23	(2.02)	-0.63	(1.17)	-3.57	(0.99)	-0.42	(0.56)
AIPW-GRF	-8.28	(1.86)	-0.04	(1.19)	-2.49	(0.92)	0.18	(0.54)

Note: ATT estimates using the IRS data. Robust standard errors are in the parentheses. The outcome is measured in thousand USD.

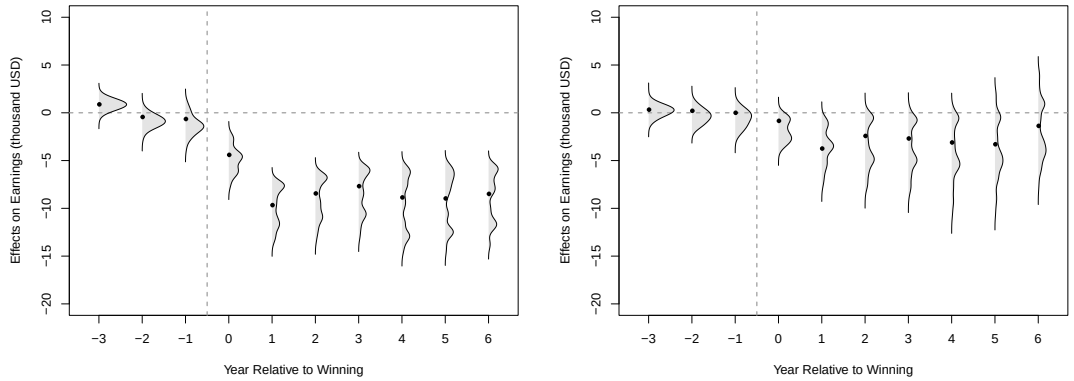
FIGURE A19. ATT and Placebo Estimates: IRS Data



Note: Subfigure (A) and (C) show the ATT estimates of big and small winners, respectively, versus non-winners on the average annual labor earnings from Year 0 to Year 6, along with their 95% confidence intervals. Subfigure (B) and (D) show the ATT estimates of big and smaller winners, respectively, versus non-winners on the placebo outcome, the average annual labor earnings from Year –3 to Year –1, along with their 95% confidence intervals. We use the same eleven estimators as before.

CATT Estimates.. Figure A20 show the CATT estimates using the IRS data.

FIGURE A20. CATT Estimates: IRS Data, Trimmed Sample

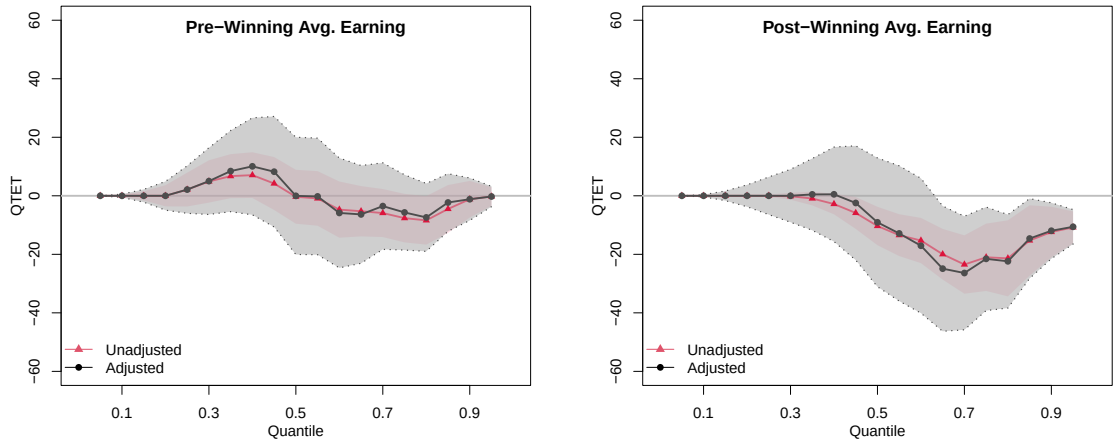


A. Big Winners vs Controls B. Small Winners vs Controls

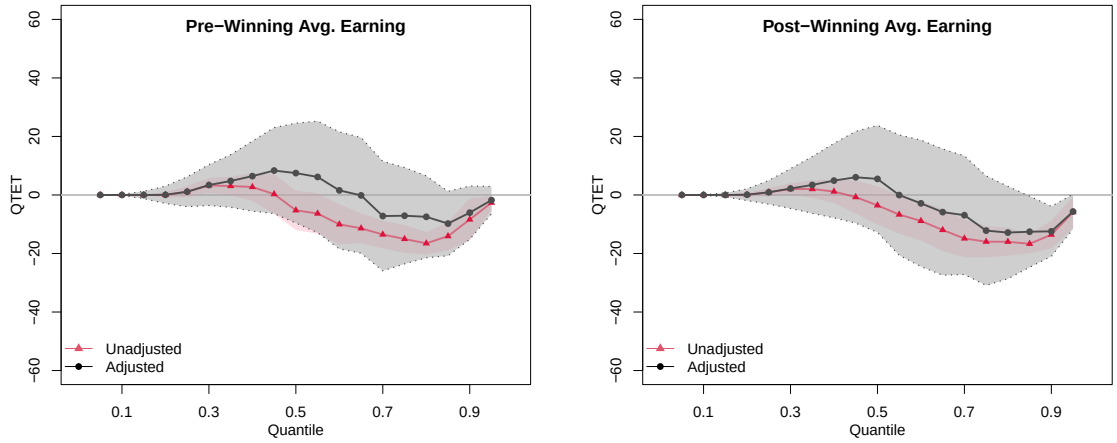
Note: Figures show the distributions of CATT estimates using the Imbens-Rubin-Sacerdote (IRS) data (with the black dots representing the corresponding ATT estimates) based on the **full samples**. We adjust for the following covariates: #tickets bought, gender, work then, age at winning, years of education, college degree, earnings 6 years before winning. The outcome variables include earnings from 5 years before winning to 6 years after winning. For each estimate, the Augmented Inverse Probability Weighting (AIPW) estimator is employed, with the Generalized Random Forest (GRF) approach for estimating nuisance parameters. The estimates for pre-winning outcomes serve as placebo tests. **Subfigure A**: "Bigger winners" versus controls (with "small winners" removed from the sample). **Subfigure B**: Small winners versus controls (with "big winners" removed from the sample).

Quantile treatment effects. Figure A21 shows the quantile treatment effects on the treated using the IRS data.

FIGURE A21. Quantile Treatment Effects: IRS Data



A. Big Winners vs Controls

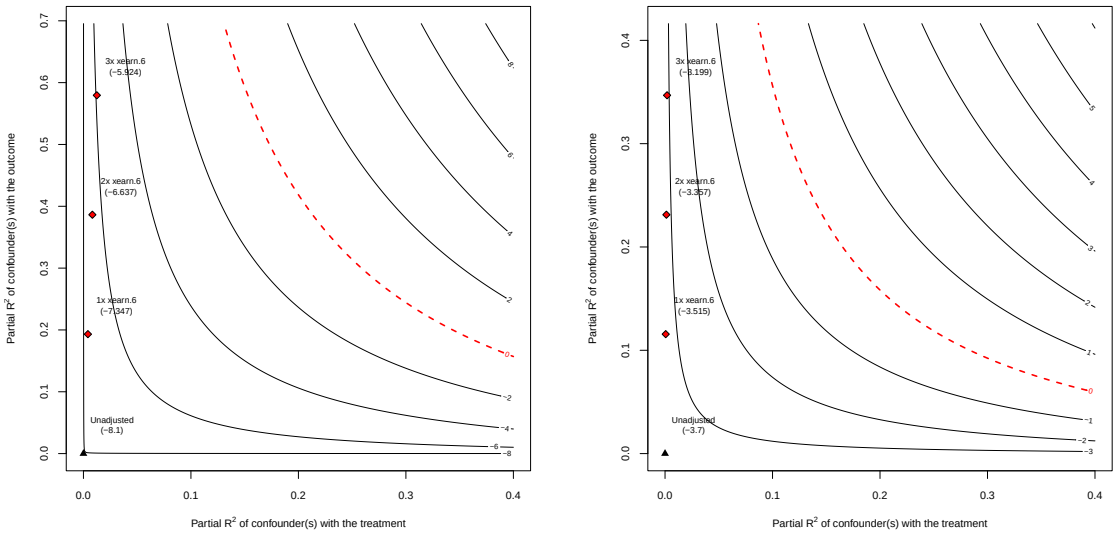


B. Small Winners vs Controls

Note: Figures show the quantile treat effects on the treated (QTET) using with or without adjusting for the covariates (in grey and pink, respectively) based on the **full samples**. Each dot corresponds to a QTET estimate at a particular quantile, while gray/pink areas represent bootstrapped 95% confidence intervals. Unadjusted models do not incorporate covariates while adjustment models use the full set of covariates (including #tickets bought, gender, work then, age at winning, years of education, college degree, and earnings 6 to 4 years before winning.) to estimate the propensity scores with a logit. Different subfigures use different samples for the nonexperimental data: **Subfigure A**: "Bigger winners" versus controls (with "small winners" removed from the sample). **Subfigure B**: Small winners versus controls (with "big winners" removed from the sample).

Sensitivity analyses. Figure A22 shows results from a sensitivity analysis using the lottery data. The estimated causal effect of big prizes is less sensitive to potential confounders than that of small prizes.

FIGURE A22. Sensitivity Analyses for the Lottery Example



A. Big Winners B. Small Winners

Note: Contour plot for the treatment effect coefficient $\hat{\tau}_{OLS}$ based on sensitivity analysis first proposed by Imbens (2003) and then modified by Cinelli and Hazlett (2020). The red dashed line indicates $\hat{\tau}_{OLS} = 0$. The benchmark covariate is earnings one year before winning the lottery. The model is a linear regression with all available long-term covariates included.

References

Susan Athey and Guido Imbens. A measure of robustness to misspecification. *The American Economic Review*, 105(5):476–480, 2015.

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

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 2020.

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

Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review, Papers and Proceedings*, 93(2):126–132, 2003.

Guido W. Imbens and Yiqing Xu. Lalonde (1986) after four decades: What lessons have we learned? *Journal of Economic Perspectives*, 2025. Forthcoming.

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.

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

Paul R Rosenbaum. *Observational Studies*. Springer, 2002.

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, 1983.