HCEO Summer School
Moscow 2017

Professor Jeffrey Smith
Department of Economics
University of Michigan (for now)
econjeff@umich.edu

Lecture 2: Experiments as Benchmarks

August 29, 2017

Outline

Definitions

LaLonde (1986) literature (almost all of the lecture)

RD literature

Structural literature

External validity

Conclusions

Key points

We can learn a lot from within study designs

Another reason to do more (carefully designed) RCTs

Much of the existing literature misinterprets what is learned

The LaLonde (1986) *AER* literature provides an example of both

Within-study designs can inform structural models as well

External validity matters here

Context

Most of the discussion resides in treatment effects land

Treatment group: experience the treatment

Comparison group: do not experience the treatment in nature

Control group: randomized individuals who would otherwise be treated

Experimental estimate: compares treatment and control groups

Non-experimental impact estimate: compares treatment and comparison groups

Non-experimental bias estimate: compares control and comparison groups)

Identification and estimation

An identification strategy is a substantive economic claim about the data generating process that allows a causal interpretation

Examples: CIA, BSA, RD

An estimation strategy is a rule for manipulating data to produce an estimate. Particular estimators are consistent under particular identifying assumptions.

Example 1: within estimator and first differences estimator under BSA
Example 2: PSM, IPW and OLS under CIA

Aside on specification / functional form

Identification = economics while estimation = econometrics

What is a within-study design?

Terminology due to Tom Cook in educational statistics

Within-study designs use experimental estimates as benchmarks to study the performance of alternative identification and estimation strategies applied to particular data in particular substantive contexts.

Why the “within”?

An extended example will make this clear.

The LaLonde (1986) literature (narrowly defined)

LaLonde (1986) *AER*: National Supported Work Demonstration (NSW)
About 1835 Google citations

Heckman and Hotz (1989) *JASA*: NSW

Heckman, Ichimura and Todd (1997) *ReStud*, Heckman, Ichimura, Smith and Todd (1998) *Econometrica*, Heckman and Smith (1999) *EJ*: National JTPA Study

Dehejia and Wahba (1999) *JASA* and (2002) *ReStat*, Dehejia (2005) *Journal of Econometrics*: NSW

Smith and Todd (2005a,b) *Journal of Econometrics*: NSW

Calónico and Smith (2017) *JoLE*: NSW

Lalonde (1986) motivation

The search for the magic bullet, the estimator that consistently solves the selection problem:

"The goal is to assess the likely ability of several econometric methods to accurately assess the economic benefits of employment and training programs" (604)

Is an "econometric method" an identification strategy or an estimator?

What about the data? The question?

This framing of the problem continues up to the present day; see e.g. Bloom, Michalopoulos and Hill (2004) *ReStat* and Hollister and Wilde (2007) *JPAM*

Lalonde (1986) basic setup

Combine the treatment group data from the NSW experiment with non-experimental comparison groups

Comparison group source one: Panel Study of Income Dynamics female heads (continuously) from 1975-1979

Comparison group source two: Current Population Survey persons in the March 1976 CPS in the labor force in 1976 with individual income < 20K and household income < 30K

Using representative samples for comparison groups was standard practice at the time – it is less so now

Three comparison groups were created from each data set using simple screens to keep in low skill individuals

Lalonde (1986) NSW experiment

The National Supported Work (NSW) Demonstration examined the impacts of an expensive treatment on four groups with labor market difficulties: long-term AFDC recipients, high school dropouts, ex-convicts and ex-addicts.

LaLonde looks at two groups: AFDC women and the men from the other three groups. Aside: why would you ever combine these groups?

Treatment group observations were randomly assigned from January 1976 to April 1977

Random assignment took place in 10 sites around the country. Not all sites served all groups. All the sites serving men were in cities, all but one of the sites serving women were in cities.

Lalonde (1986) identification strategies and estimators

Identification: Selection on observed variables

- Demographics and education
- Demographics, education and pre-program earnings

Identification: Bias stability a.k.a. common trends

Identification: Normal selection model

- No exclusion restriction (i.e. functional form only)
- Dubious exclusion restrictions (more on this later)

All linear parametric estimators – it was the early 1980s.

- Heckman two-step estimator for the normal model

Lalonde (1986) variables

Covariates: age, Black and Hispanic indicators, years of schooling, an indicator for married, and a high school completion indicator (and that is all!)

Outcome variable: Real earnings from NSW survey (treatment group), SSA earnings records (CPS comparison group), PSID survey (PSID comparison group)

Dependent variable: real earnings in 1979

Lagged dependent variable: real earnings in 1975

Exclusion restriction variables: urban residence (!), employment status in 1976 (!), AFDC status in 1975 (!), and number of children (!)

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.

The researchers who evaluated these federally sponsored programs devised both experimental and nonexperimental procedures to estimate the training effect, because they recognized that the difference between the trainees' pre- and post-training earnings was a poor estimate of the training effect. In a dynamic economy, the trainees' earnings may grow even without an effective program. The goal of these program evaluations is to estimate the earnings of the trainees had they not participated in the program. Researchers using experimental data take the earnings of the control group members to be an estimate of the trainees' earnings without the program. Without experimental data, researchers estimate the earnings of the trainees by using the regression-adjusted earnings of a comparison group drawn from the population. This adjustment takes into account that the observable characteristics of the trainees and the comparison group members differ, and their unobservable characteristics may differ as well.

Any nonexperimental evaluation of a training program must explicitly account for these differences in a model describing the observable determinants of earnings and the process by which the trainees are selected into the program. However, unlike in an experimental evaluation, the nonexperimental estimates of the training effect depend crucially on the way that the earnings and participation equations are specified. If the econometric model is specified correctly, the nonexperimental estimates should be the same (within sampling error) as the training effect generated from the experimental data, but if there is a significant difference between the nonexperimental and the experi-

ployment and Training Act–1973, and the Job Training Partnership Act–1982.

Lalonde (1986) results

The non-experimental impact estimates vary widely across identification strategies

The non-experimental impact estimates vary widely across comparison groups

Limited specification tests combined with a priori reasoning do not rule out all of the poorly-performing estimators

Bivariate normal model results are wrong for the reasons already described plus problems with choice-based sampling.

Lalonde (1986) conclusions

"… policymakers should be aware that the available non-experimental evaluations of employment and training programs may contain large and unknown biases resulting from specification errors." (617)

Absolutely!

But, this paper was widely interpreted to mean that only experiments could provide credible impact estimates for active labor market policies.

It directly resulted in the choice of an experimental design for the National Job Training Partnership Act Study. It indirectly helped spawn the "credibility revolution"

Thought question: why doesn't this paper condemn all empirical work in applied microeconomics?

Lalonde (1986) alternative reading

The data are much (all?) of the problem, not the methods, particularly for the men.

Why would you expect the handful of covariates here to solve the selection problem?

No measures related to crime for the ex-convicts or to substance use for the ex-addicts

Lagged annual earnings not well aligned with the time of treatment and measured and aligned differently for the treated and untreated units.

And we blame the methods?

Heckman and Hotz (1989) basic setup

These two, along with LaLonde, were my dissertation committee at Chicago

HH (1989) *JASA* use administrative earnings data for the NSW and CPS samples

The data are grouped, which rules out non-linear estimators

The NSW sample gets a lot bigger as attrition is no longer an issue

Examine CIA/OLS, BSA/fixed effects with different “before” periods, random growth with different “before” periods

Heckman and Hotz (1989) specification tests

Systematically apply specification tests

The "pre-program" test

Apply the estimator to outcomes before treatment and test for a zero impact

Tests of over-identifying restrictions

For FE and random growth models test for zero coefficients on lags of Y that should not show up as regressors if the model is correct

Heckman and Hotz (1989) conclusions

Find that they can reject all models that differ in both sign and statistical significance from the experimental results (but point estimates vary among models not rejected)

Heckman has since more or less come out against pre-program tests.

I would argue that they have proven useful in the literature at removing really bad estimates.

There is more research to be done here.

HIT (1997), HIST (1998), Heckman and Smith (1999) basic setup

NJS is a random assignment evaluation of parts of the Job Training Partnership Act (JTPA) program

Non-random set of 16 of 600 “service delivery areas”

At four of the 16 sites “ideal” comparison group data on Eligible Non-Participants (ENPs) collected:

Same survey instrument
Same local labor markets
All comparison group members screened for eligibility
Detailed employment, earnings and labor force histories

The same data were collected for control group members at these sites

Heckman et al. (1997, 1998, 1999) analysis

Also construct a comparison group from the Survey of Income and Program Participation (SIPP)

Assume the CIA or BSA and apply various matching methods (nearest neighbor, kernel, and local linear) to the ENPs and controls at the four sites

Larger sample sizes than NSW but still not that large, especially for youth

Heckman et al. (1997, 1998, 1999) findings

Measuring the dependent variable in the same way for treated and untreated units matters

The SIPP and ENP survey measures are very different conditional on *X* and the administrative and survey measures differ for the ENPs

Drawing treated and untreated observations from the same local labor markets matters

The SIPP never performs very well, nor do cross-site comparisons using the controls and ENPs

Heckman et al. (1997, 1998, 1999) more findings

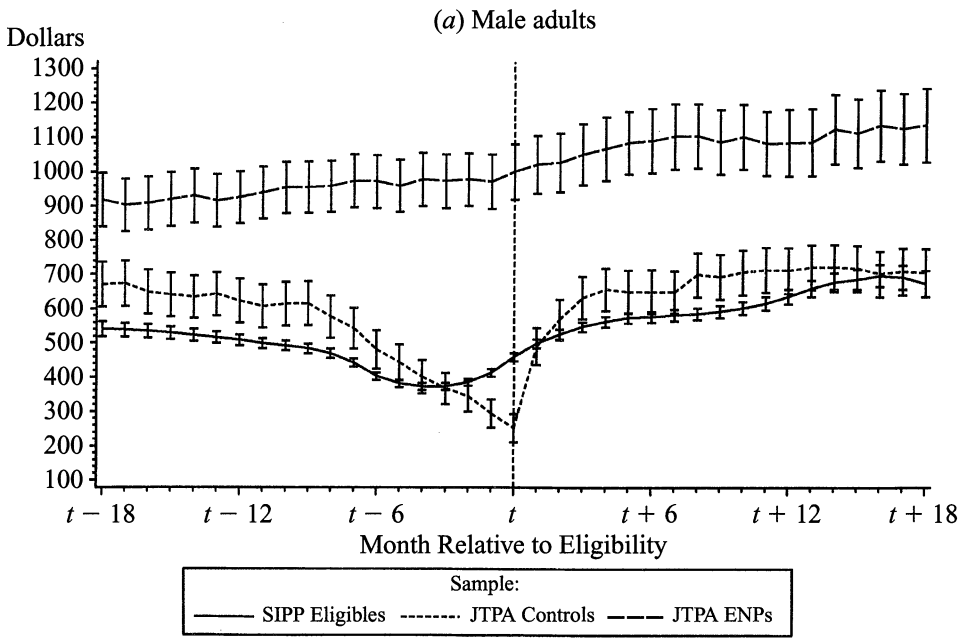
What you condition on matters for the plausibility of the CIA

Conditioning on detailed labor force status histories matters a lot for adult males at reducing the bias

Longitudinal estimators perform very poorly due to BSA failure

See Heckman and Smith (1999) *EJ*

They are undone by the pre-program dip combined with post-program earnings growth for controls relative to ENPs



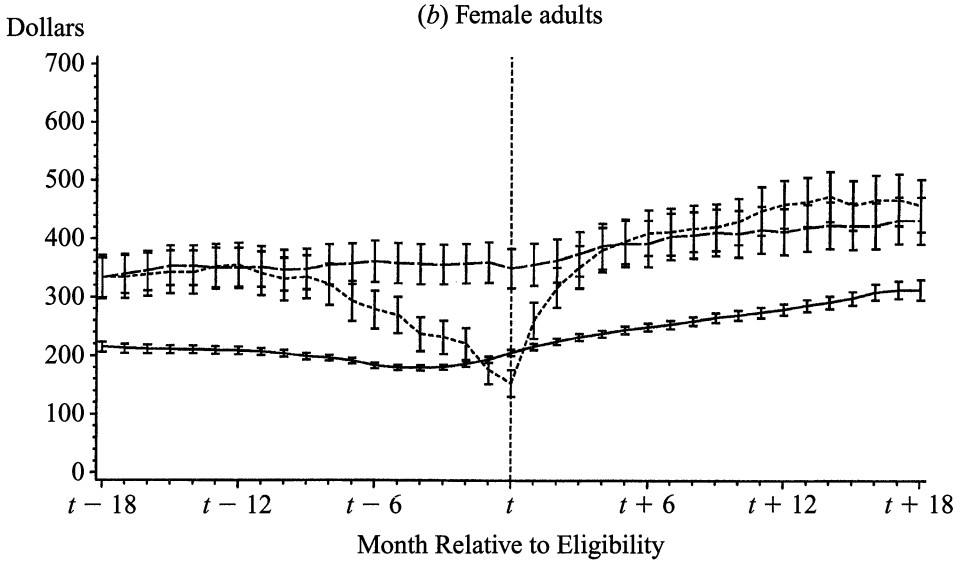


Fig. 1. *Mean Self-Reported Monthly Earnings: SIPP Eligibles and JTPA Controls and ENPs*
Notes: SIPP uses all JTPA–eligible person–month observations of respondents present in both the first and last months of the panel. Controls are randomised–out participants from the National JTPA Study. Observations based on quasi-rectangular sample. ENPS are JTPA–eligible non-participants at the same sites as the controls from the National JTPA Study. Observations based on quasi-rectangular sample. Standard error bars +/− 2 standard errors of the means.

Dehejia and Wahba (1999, 2002) basic setup

Assume the CIA and apply propensity score matching to a subset of LaLonde's male sample

Strong a priori case for matching rather than parametric linear model in this context due to support issues

Implicitly a paper about estimators and not about identification strategies

The sad story of the tapes

Foreshadow Calónico and Smith (2017)!

Dehejia and Wahba (1999, 2002) estimation

Matching variables: same as LaLonde plus real earnings in 1975, real earnings in "1974" and indicators for zero earnings in "1974" and 1975

Motivation: want to condition on past outcomes! See Ashenfelter (1978) *ReStat*, Card and Sullivan (1988) *Econometrica* etc.

Choose exact specification using balancing tests.

But are a couple of years of misaligned earnings measured differently for the treated and untreated units enough?

Apply single nearest neighbor matching with replacement, inverse probability weighting (in Dehejia's dissertation) and stratification.

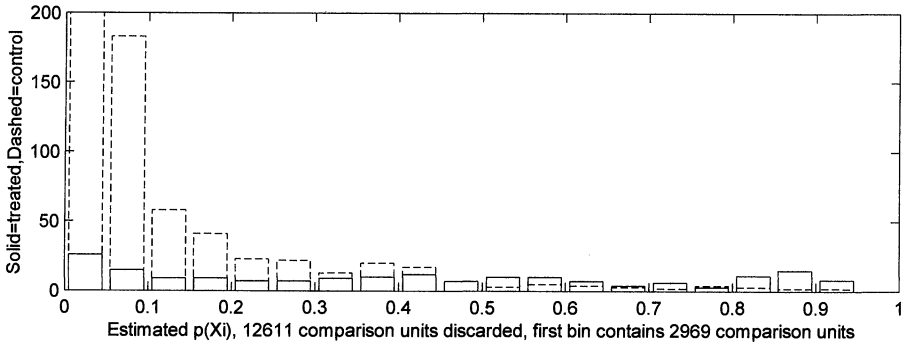


Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12,611 CPS units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 2,969 CPS units. There is minimal overlap between the two groups, but the overlap is greater than in Figure 1; only one bin (.45–.5) contains no comparison units, and there are 35 treated and 7 comparison units with an estimated propensity score greater than .8.

treatment group (although the treatment impact still could be estimated in the range of overlap). With limited overlap, we can proceed cautiously with estimation. Because in our application we have the benchmark experimental estimate, we are able to evaluate the accuracy of the estimates. Even in the absence of an experimental estimate, we show in Section 5 that the use of multiple comparison groups provides another means of evaluating the estimates.

We use stratification and matching on the propensity score to group the treatment units with the small number of comparison units whose estimated propensity scores are greater than the minimum—or less than the maximum—propensity score for treatment units. We estimate the treatment effect by summing the within-stratum difference in means between the treatment and comparison observations (of earnings in 1978), where the sum is weighted by the number of treated observations within each stratum [Table 3, column (4)]. An alternative is a within-block regression, again taking a weighted sum over the strata [Table 3, column (5)]. When the covariates are well balanced, such a regression should have little effect, but it can help eliminate the remaining within-block differences. Likewise for matching, we can estimate a difference in means between the treatment and matched comparison groups for earnings in 1978 [column (7)], and also perform a regression of 1978 earnings on covariates [column (8)].

Table 3 presents the results. For the PSID sample, the stratification estimate is \$1,608 and the matching estimate is \$1,691, compared to the benchmark randomized-experiment estimate of \$1,794. The estimates from a difference in means and regression on the full sample are −\$15,205 and \$731. In columns (5) and (8), controlling for covariates has little impact on the stratification and

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		*NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score*					
			Quadratic in score[b]	*Stratifying on the score*			*Matching on the score*	
	(1) Unadjusted	*(2) Adjusted*[a]	*(3)*	*(4) Unadjusted*	*(5) Adjusted*	*(6) Observations*[c]	*(7) Unadjusted*	*(8) Adjusted*[d]
NSW	1,794	1,672						
	(633)	(638)						
PSID-1[e]	−15,205	731	294	1,608	1,494	1,255	1,691	1,473
	(1,154)	(886)	(1,389)	(1,571)	(1,581)		(2,209)	(809)
PSID-2[f]	−3,647	683	496	2,220	2,235	389	1,455	1,480
	(959)	(1,028)	(1,193)	(1,768)	(1,793)		(2,303)	(808)
PSID-3[f]	1,069	825	647	2,321	1,870	247	2,120	1,549
	(899)	(1,104)	(1,383)	(1,994)	(2,002)		(2,335)	(826)
CPS-1[g]	−8,498	972	1,117	1,713	1,774	4,117	1,582	1,616
	(712)	(550)	(747)	(1,115)	(1,152)		(1,069)	(751)
CPS-2[g]	−3,822	790	505	1,543	1,622	1,493	1,788	1,563
	(670)	(658)	(847)	(1,461)	(1,346)		(1,205)	(753)
CPS-3[g]	−635	1,326	556	1,252	2,219	514	587	662
	(657)	(798)	(951)	(1,617)	(2,082)		(1,496)	(776)

[a] Least squares regression: RE78 on a constant, a treatment indicator, age, age^2, education, no degree, black, Hispanic, RE74, RE75.

[b] Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).

[c] Number of observations refers to the actual number of comparison and treatment units used for (3)–(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

[d] Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation [same covariates as (a)]. Propensity scores are estimated using the logistic model, with specifications as follows:

[e] PSID-1: Prob (T_i = 1) = F(age, age^2, education, education^2, married, no degree, black, Hispanic, RE74, RE75, RE74^2, RE75^2, u74*black).

[f] PSID-2 and PSID-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE74^2, RE75, RE75^2, u74, u75).

[g] CPS-1, CPS-2, and CPS-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE75, u74, u75, education*RE74, age^3).

Dehejia and Wahba (1999, 2002) conclusions

"The methods we suggest are not relevant in all situations. There may be important unobservable covariates, for which the propensity score method cannot account. However, rather than giving up, or relying on assumptions about the unobserved variables, there is substantial reward in exploring first the information contained in the variables that *are* observed."

This is all quite correct.

But: unobservable or unobserved?

Dehejia and Wahba (1999, 2002) conclusions (continued)

The literature reads this paper to mean that “matching works” even in weak data contexts

The story of the zoning paper

This common interpretation of the DW (1999, 2002) papers represents an incorrect answer to an ill-posed question.

We already know when matching works. It works when the conditioning variables satisfy the conditional independence assumption in the substantive context at hand.

This is the key misinterpretation in the literature: it informs us about what variables lead the CIA to hold in particular substantive contexts, not about the success of candidate magic bullet estimators

Smith and Todd (2005) NSW

Concern: given the Heckman et al. findings in the NJS, why should matching ever work with the NSW data, given the implausibility of the CIA, at least for the men?

Recall: dependent variable measured differently by treatment status, different local labor markets, very limited conditioning variables

Replicates Dehejia and Wahba (1999, 2002)

Their results are correct as stated; if you do what they did, you get what they got – that's not true of every paper!

Performs numerous sensitivity analyses; we focus on just one for reasons of time and quickly review the others

TABLE 2
Dehejia and Wahba (1999,2002) Sample Composition
Month of Random Assignment and
Earnings 13-24 Months Before Random Assignment
Number in Cell, Row Percentage and Overall Percentage
Shaded Area Indicates DW Sample

Month of Random Assignment	Zero Earnings in Months 13-24 Before RA	Non-Zero Earnings in Months 13-24 Before RA
August 1977	7 46.67 0.97	8 53.33 1.11
July 1977	24 41.38 3.32	34 58.62 4.71
January 1977	6 50.00 0.83	6 50.00 0.83
December 1976	53 36.81 7.34	91 63.19 12.60
November 1976	43 40.57 5.96	63 59.43 12.60
October 1976	63 45.99 8.73	74 54.01 10.25
April 1976	37 59.68 5.12	25 40.32 3.46
March 1976	35 47.30 4.85	39 52.70 5.40
February 1976	33 49.25 4.57	34 50.75 4.71
January 1976	26 55.32 3.60	21 44.68 2.91

Smith and Todd (2005) analysis

Restricting attention to the early random assignment sample does away with the low bias estimates very quickly.

Should it? What about the alignment concerns?

This result is insensitive to alternative matching schemes

The same is true if the entire LaLonde (1986) sample of men is used

See also Diamond and Sekhon (2013) *ReStat* who apply Genmatch to all three samples.

TABLE 5A
Bias Associated with Alternative Cross-Sectional Matching Estimators
Comparison Group: CPS Male Sample
Dependent Variable: Real Earnings in 1978
(Bootstrap Standard Errors in Parentheses; Trimming Level for Common Support is Two Percent)

Sample and Propensity Score Model	(1) Mean Diff.	(2) 1 Nearest Neighbor Without Common Support	(3) 10 Nearest Neighbor Estimator without Common Support	(4) 1 Nearest Neighbor Estimator with Common Support	(5) 10 Nearest Neighbor Estimator with Common Support	(6) Local Linear Matching (bw = 1.0)	(7) Local Linear Matching (bw =4.0)	(8) Local Linear Regression Adjusted Matching[a] (bw =1.0)	(9) Local Linear Regression Adjusted Matching (bw =4.0)
Lalonde Sample with DW Prop. Score Model	-9757	-555	-270	-838	-1299	-1380	-1431	-1406	-1329
	(255)	(596)	(493)	(628)	(529)	(437)	(441)	(490)	(441)
as % of $886 impact	-1101%	-63%	-30%	-95%	-147%	-156%	-162%	-159%	-150%
	(29)	(67)	(56)	(71)	(60)	(49)	(50)	(55)	(50)
DW Sample with DW Prop. Score Model	-10291	407	-5	-27	-261	-88	-67	-127	-96
	(306)	(698)	(672)	(723)	(593)	(630)	(611)	(709)	(643)
as % of $1794 impact	-574%	23%	-0.3%	-1.5%	-15%	-5%	-4%	-5%	-7%
	(17)	(39)	(37)	(40)	(33)	(35)	(34)	(40)	(36)
Early RA sample with DW Prop. Score Model	-11101	-7781	-3632	-5417	-2396	-3427	-2191	-3065	-3391
	(461)	(1245)	(1354)	(1407)	(1152)	(1927)	(1069)	(3890)	(1124)
as % of $2748 impact	-404%	-283%	-132%	-197%	-87%	-125%	-80%	-112%	-123%
	(17)	(45)	(49)	(51)	(42)	(70)	(39)	(142)	(41)
Lalonde Sample with Lalonde Prop. Score Model	-10227	-3602	-2122	-3586	-2342	-3562	-2708	-3435	-2362
	(296)	(1459)	(1299)	(1407)	(1165)	(3969)	(1174)	(4207)	(1178)
as % of $886 impact	-1154%	-406%	-240%	405%	264%	402%	306%	388%	-266%
	(33)	(165)	(147)	(159)	(131)	(448)	(133)	(474)	(133)

a. Regression adjustment includes race and ethnicity, age categories, education categories and married.

Smith and Todd (2005) additional analyses

Small changes in conditioning variables (compare DW scores with LaLonde participation model less urban residence):

Changing the propensity score specification matters a lot

Alternative matching estimators (nearest neighbor, multiple nearest neighbor, local linear, and regression-adjusted local linear):

No systematic differences

Smith and Todd (2005) even more analyses

Ties (vary random number generator seed):

Matters in the CPS sample due to several ties, small sample size, and large variance in earnings among observations with identical scores

Imposing or not imposing the common support condition and how to impose it:

Makes no difference

Smith and Todd (2005) yet more analyses

Data used to estimate the propensity score (control group, treatment group or both):

The bias can differ by hundreds of dollars depending on which observations are used to estimate the scores (all of them consistent!)

Bias estimates rather than impact estimates:

The bias can differ by several hundred dollars depending on whether the treatment group or the control group is used to estimate it

Alternative balancing tests:

DW scores fail many other tests in the literature (but …)

The point of reviewing the NSW papers

The point is not to bash LaLonde or Dehejia and Wahba! They are smart fellows.

The point is to learn the right lessons from this evidence

The sample sizes are really small. For the men: LaLonde (T 297, C 425), DW (T 185, C260), Early RA (T 108, C 142)

The comparison groups are not very comparable for the men

The results are pretty sensitive no matter what you do (did I mention that the sample is really small and we are doing semi-parametrics?)

Is this the best data set to use to test estimators? Probably not!

Calónico and Smith (2017)

Recreate the AFDC women samples from LaLonde (1986)

NSW: no problem

PSID: issues, but is it the PSID version or something else?

Do with them (much of) what LaLonde (1986) did, what DW (1999, 2002) did and what ST (2005a,b) did …

A deeper kind of replication than just re-running the author's program on the author's analysis file

Both a replication and an extension in the nomenclature of the Clemens (2015) *JES*

Calónico and Smith (2017) results

Conditioning the comparison group on any AFDC receipt in 1975 (and nothing else) reduces bias substantially to relatively modest levels

The early RA sample performs better rather than worse

Implication: alignment matters for the women

Normalized IPW reduces variance a lot without a bias cost

Calónico and Smith (2017) results (continued)

The women pose a generally less difficult selection problem:

- Less distinct propensity score distributions, especially with PSID-2
- Matching does not consistently reduce the estimated bias relative to the parametric linear model

Difference-in-differences matching does not reduce bias very much.

What about geographic mismatch and different earnings measures?

- The time-invariant bias found for the men must be something specific to the men.

RD within-study designs

Black, Galdo and Smith (under construction, for a very long time)

Buddelmeyer and Skoufias (2004) IZA

See also the recent unpublished survey by Tom Cook and co-authors

Key issue here is to match the experimental estimand to the RD estimand

Are these exercises inherently less interesting?

Within-study designs and structural estimation

Todd and Wolpin (2006) *AER*

Alan Griffith (2017) job market paper

Lise, Seitz and Smith (2004) NBER

Partial equilibrium versus general equilibrium

Key tradeoff: use experimental variation for testing or identification?

External validity

Shadish, Clark and Steiner (2008) *JASA*

Randomly assign undergraduates in a large psychology class to be either (1) randomly assigned to mathematics training or vocabulary training or (2) allowed to self-select into either mathematics training or vocabulary training.

Collect a thoughtful set of baseline covariates related to training choice and outcomes, which consist of scores on mathematics and vocabulary tests

External validity (continued)

CIA does well here whether implement as OLS estimation of a parametric linear model or matching or weighting

So what?

Refer back to the question about generalizing LaLonde (1986) to other substantive domains

This study strikes me as cool in its design but of low value

Economics needs a better, empirically grounded theory of external validity

Aside: external validity is not just an issue with RCTs

Other recent literature using experimental benchmarks

Training programs

Fraker and Maynard (1987) *JHR* NSW
Bell, Orr, Blomquist and Cain (1995) Upjohn book: AFDC Homemaker Home Health Aide Demonstration

Welfare to work programs and geographic selection

Friedlander and Robins (1995) *AER* Various welfare-to-work programs
Bloom, Michaelopoulos and Hill (2004) *ReStat*: NEWWS

Education programs

Agodini and Dynarski (2004) *ReStat*: dropout prevention programs
Hollister and Wilde (2007) *JPAM*: Project STAR

Other recent literature with experimental benchmarks - continued

Development

Diaz and Handa (2006) *JHR* PROGRESA
[Diaz is my student from Maryland]

Outside economics

Heinsman and Shadish (1996) *Psychological Methods*
Lipsey and Wilson (1993) *American Psychologist*

Conclusions (narrow)

LaLonde (1986) does not show that non-experimental estimators do not work.

Dehejia and Wahba (1999, 2002) do not show that propensity score matching “works”

Specification tests such as the “pre-program” test are worthy of further study

Replication is useful and can be carried out peacefully

Conclusions (broad)

We can learn a lot about non-experimental identification strategies and estimators using experiments as benchmarks. This is another reason for doing more (thoughtful) RCTs.

The correct lessons:

There is no magic bullet – no estimator that always solves the selection problem

The question is not “which estimator works.” Instead, we want to know the mapping from data, parameter of interest and institutional context to estimator choice

Clever econometrics will usually lose out to bad data