The Women of the National Supported Work Demonstration

Sebastian Calónico, *University of Miami*

Jeffrey Smith, *University of Michigan, NBER, IZA, and CESifo*

This paper re-creates three of the samples from LaLonde's famous 1986 paper that began the literature on "within-study designs" that uses experiments as benchmarks against which to assess the performance of nonexperimental identification strategies. In particular, we recreate the experimental data for the target group of women on welfare from the National Supported Work (NSW) Demonstration and two of the corresponding comparison groups drawn from the Panel Study of Income Dynamics (PSID). The loss of these data resulted in the (sizable) subsequent literature devoting its attention solely to the NSW men. In addition to repeating LaLonde's analyses on our recreations of his files for the AFDC women, we apply (many of) the estimators from later papers by Dehejia and Wahba and by Smith and Todd to these data. Our findings support the general view in the literature that women on welfare pose a less difficult selection problem when evaluating employment and training programs. They also call into question the generalizability of some of the broad conclusions that Dehejia and Wahba and Smith and Todd draw from their analyses of the NSW men.

I. Introduction

Within-study comparisons using experiments as benchmarks against which to judge the performance of nonexperimental identification strategies

We thank participants at the conference in honor of Robert LaLonde held at the Chicago Federal Reserve in April 2015 for helpful comments, particularly our dis-

[*Journal of Labor Economics*, 2017, vol. 35, no. S1]
© 2017 by The University of Chicago. All rights reserved. 0734-306X/2017/35S1-0010$10.00
Submitted August 14, 2015; Accepted March 31, 2017

applied using particular data sets in specific programmatic contexts have played a major role in the development of economists' thinking about how best to evaluate active labor market programs and how best to undertake empirical work aimed at estimating causal effects more generally. The within-study comparison literature begins with LaLonde's (1986) widely cited and justly famous study that combines the experimental data from the National Supported Work (NSW) Demonstration with nonexperimental data drawn from two large social science datasets.

In addition to the long trail of studies that reuse the data on men from LaLonde (1986), the within-study comparison literature includes a number of papers based on the data from the National Job Training Partnership Act Study (NJS), which, inspired by LaLonde (1986) and the ensuing discussion, incorporated a within-study comparison component into the original design and data collection. In recent years, many experimental evaluations have formed the basis of such comparisons, including the Tennessee STAR class size experiment in Hollister and Wilde (2007), the Canadian Self-Sufficiency Program experiment in Lise, Seitz, and Smith (2004), and the Progresa conditional cash transfer program experiment in Mexico in Todd and Wolpin (2006). This research program has generated much knowledge about what nonexperimental identification strategies work with what data in particular institutional contexts, as well as about the performance of structural models. As a result, it helps researchers predict when particular nonexperimental strategies will (and will not) provide estimates close to those an experiment would have provided, and it provides guidance on what data elements matter for future nonexperimental evaluation efforts.

Because LaLonde's (1986) analysis file for men survived but that for women did not, the (vast) subsequent literature that builds on his study analyzes only the men, using his analysis file as a base. In light of this startling lacuna in the literature, we return to the raw data sets underlying LaLonde's

cussant Jens Ludwig. We also received valuable comments at the Danish Microeconometric Network meetings in Copenhagen (2009), the 3rd Joint IZA/IFAU Conference on Labor Market Policy Evaluation in Uppsala (2010), the CESifo education group meetings in Munich (2015), the CAFÉ conference in Denmark (2015), and seminar presentations at Michigan, Northwestern, and Notre Dame. Louis-Pierre Lepage provided valuable research assistance with enthusiasm and interest. The reasoning in this paper has benefited from many general discussions over the years with James Heckman, Dan Black, Michael Lechner, Petra Todd, Barbara Sianesi, and Jessica Goldberg. We thank Barbara Sianesi and Frank Stafford for especially careful readings of earlier versions, Charlie Brown for useful discussions about the PSID, and Tom Cook for a helpful e-mail exchange on the intellectual history of within-study comparisons. Much as we might like to blame any errors on anonymous passers-by, we must instead retain responsibility for any that might appear. Contact the corresponding author, Jeffrey Smith, at econjeff@umich.edu. Information concerning access to the data used in this article is available as supplementary material online.

(1986) work and do our best to recreate his analysis files for women. We then use our versions of his analysis files to replicate his analysis of the Aid to Families with Dependent Children (AFDC) women target group in the NSW demonstration.[1] We also repeat (a subset of) the analyses in Dehejia and Wahba (1999, 2002) and Smith and Todd (2005a, 2005b) using our data on the AFDC women. The Dehejia and Wahba (1999, 2002) papers played a crucial role in introducing matching estimators from the applied statistics literature into the applied economics literature. Smith and Todd (2005a, 2005b) helped to popularize the difference-in-differences matching estimators introduced in Heckman et al. (1998). More importantly, they curbed the enthusiasm of the empirical literature for propensity score matching engendered by overly optimistic readings of Dehejia and Wahba (1999, 2002) by revealing the sensitivity of their cheery findings. Repeating the Dehejia-Wahba and Smith-Todd analyses using the AFDC women reveals whether the conclusions they draw generalize to another target group that experienced the same program at the same time and for whom we have the same data.

In the language of Clemens (2015), we replicate the AFDC women component of LaLonde (1986) while extending the analyses of Dehejia and Wahba (1999, 2002) and Smith and Todd (2005a, 2005b) to an additional target group. The conceptual "big tent" of replication includes many types of analyses, ranging from simply rerunning code provided by authors on analysis files provided by authors to see if the same answers emerge to recreating an entire analysis from scratch starting with the original raw data sets. We undertake the latter, both because we think it provides a unique and difficult test of the original analysis and because the absence of LaLonde's (1986) analysis files and his code makes it impossible to do otherwise.[2]

We learn five main lessons from this exercise. First, though we do not exactly replicate LaLonde's (1986) analysis files, our best attempt at recreating them yields the same qualitative conclusions. Second, the literature in the United States contains hints that women on welfare represent a less challenging evaluation problem than do disadvantaged men; see, for example, the findings in Friedlander and Robins (1995). The women typically have a less dramatic pre-program earnings dip and appear to select more randomly (i.e., in a way less correlated with the unobserved component of the untreated outcome) into programs conditional on eligibility. Our findings of substantively low biases for the most plausible combinations of comparison groups and identification strategies and of small differences between the es-

[1] AFDC is the predecessor of the current Temporary Assistance for Needy Families (TANF) program. It provided cash assistance to (mostly) single mothers, as well as categorical eligibility for health insurance via Medicaid.

[2] See, e.g., Hamermesh (2007, 2016) and Duvendack, Palmer-Jones, and Reed (2015) for more on replication in economics.

timates generated by parametric linear models and semi-parametric matching and weighting estimators comport with this view. Third, our results support the common-sense notion that comparison groups, to the extent possible, should comprise only individuals eligible for the treatment under study. Fourth, when nonrandom selection due to the pre-program earnings dip matters less, issues of temporal alignment between the participation choice and the measurement of time-varying conditioning variables matter relatively more. Fifth, and finally, the limited differences we find between cross-sectional and difference-in-differences identification strategies for women imply that the substantial differences that Smith and Todd (2005a) find for the men must result primarily from factors specific to the men, such as residual selection on unobserved variables, rather than from the common factors, such as differences in the measurement of the dependent variable or geographic mismatch, that they emphasize.

The remainder of the paper leads the reader down the following path: Section II describes the now ancient National Supported Work Demonstration that provides our experimental data. Section III details the conceptual framework for within-study comparisons. Section IV describes the LaLonde (1986) study in general and our replication of his analysis of the long-term AFDC women target group. Section V presents our extension of Dehejia and Wahba (1999, 2002), and Section VI does the same for our extension of Smith and Todd (2005a, 2005b). Section VII concludes.

II. The National Supported Work Demonstration

The National Supported Work (NSW) Demonstration was a transitional, subsidized work experience program that operated between 1974 and 1979 at 15 locations throughout the United States. Four target groups were selected for inclusion in the program: female long-term AFDC recipients, former drug addicts, ex-offenders, and young school dropouts. The program provided trainees with work in a sheltered training environment and then assisted them in finding regular jobs. In providing these services, Supported Work spent far more per participant—around $14,000 in direct program operating costs in 1997 dollars—than typically spent under other programs, such as the Workforce Innovation and Opportunity Act (WIOA).[3]

The NSW eligibility criteria sought to identify individuals with strong barriers to finding a job. The main criteria were (1) the person must have been currently unemployed (defined as having worked no more than 40 hours in the 4 weeks preceding the time of selection into the program), and (2) the person must have spent no more than 3 months on one regular job of at least 20 hours per week during the preceding 6 months. For the AFDC target group, additional criteria applied: (3) no child age less than 6 years old and

[3] See, e.g., the discussions around table 18 of Heckman, LaLonde, and Smith (1999) and the references therein.

(4) on AFDC for at least 30 of the last 36 months.[4] The program was voluntary, and participants constituted a tiny fraction of the eligible population, implying substantial scope for nonrandom selection into the program.

The NSW demonstration operated as a randomized experiment in 10 of its 15 sites. Along with the Negative Income Tax (NIT) experiments, the NSW represented one of the first major social experiments in the United States (and, indeed, in the world). The overall experimental sample includes 6,616 treatment and control observations, for which data were gathered through a retrospective baseline interview and multiple follow-up surveys.[5] As noted in Heckman et al. (2000, table 1), the NSW experiment stands out for its low rates of treatment group dropout and control group substitution.[6] As such, the experimental impact estimates, which formally provide estimates of the impact of the "intention to treat" (i.e., of the offer of NSW), also closely approximate the ATET ("average treatment effect on the treated").[7]

Seven of the experimental sites served AFDC women, with random assignment at one or more of these sites in operation from February 1976 through August 1977. On average, women in the treatment group spent 9.0 months in Supported Work.[8] Figure 1 graphs the mean earnings of the experimental treatment and control groups for the AFDC women (along with the comparison groups, to which we turn later on). The graph reveals very strong impacts during 1976 and 1977, when most treatment group members held Supported Work jobs, and smaller, but still statistically and substantively significant earnings impacts in later years. Couch (1992) provides long-term impact estimates for LaLonde's male and female samples using administrative data. He finds that the impacts for the AFDC women persist through 1986.

III. Within-Study Comparisons

The deepest contribution of LaLonde (1986) consists of his introduction of what the literature has come to call "design experiments," or "within-

[4] Younger readers may not realize that back in the dinosaur days of the 1970s, government programs typically did not push unmarried mothers of children below school age into work.

[5] Response rates were an issue; see the discussion in LaLonde (1986).

[6] We suspect that "randomization bias" poses a minor issue in the NSW Demonstration as well due to the volunteer aspect of the program as well as the expensive services on offer. Also, Hollister (1984, 35) notes that "there was very little resentment of the random assignment process on the part of the enrollees themselves" in the context of his colorful description of how much the NSW program operators hated it. See Sianesi (2017) for a thorough discussion of (and some evidence on) randomization bias.

[7] See Hollister, Kemper, and Maynard (1984) for a book-length overview of the NSW Demonstration and Kemper, Long, and Thornton (1981) for the full cost-benefit analysis.

[8] See Hollister and Maynard (1984, table 4.4) and the surrounding text for more on the participation experience.

Table 1
Descriptive Statistics for Aid to Families with Dependent Children (AFDC) Experimental and Comparison Group Samples

	LaLonde Data		Calónico-Smith Data			
					Comparison Group	
Variable	Treatments	Controls	Treatments	Controls	PSID-1	PSID-2
Age	33.37	33.63	33.33	33.46	37.07	34.54
	(7.43)	(7.18)	(7.52)	(7.57)	(10.57)	(9.34)
Years of school	10.30	10.27	10.27	10.27	11.30	10.49
	(1.92)	(2.00)	(2.03)	(2.00)	(2.77)	(2.13)
Proportion high school dropouts	.70	.69	.70	.69	.45	.59
	(.46)	(.46)	(.46)	(.46)	(.50)	(.49)
Proportion married	.02	.04	.02	.04	.02	.01
	(.15)	(.20)	(.15)	(.21)	(.14)	(.10)
Proportion black	.84	.82	.84	.82	.65	.86
	(.37)	(.39)	(.37)	(.39)	(.48)	(.35)
Proportion Hispanic	.12	.13	.12	.13	.02	.02
	(.32)	(.33)	(.32)	(.33)	(.12)	(.15)
Month of assignment (January 1978 = 0)	−12.26	−12.30	−12.23	−12.26		
	(4.3)	(4.23)	(4.39)	(4.4)		
No. of observations	800	802	800	802	648	182

NOTE.—Standard deviations appear in parentheses.

study" comparisons.[9] Such comparisons use (most often) experimental evaluations as benchmarks against which to measure the performance of various nonexperimental estimators applied to data on nonexperimental comparison groups in particular programmatic contexts.

To formalize the notion of a within-study design, consider the standard potential outcomes framework, wherein Y_{1i} denotes the outcome with treatment for unit i and Y_{0i} the outcome without treatment for the same unit. Let $D_i \in \{0, 1\}$ indicate treatment choice in the absence of random assignment. In observational data, the observed outcome has a simple switching regression representation as $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$. We assume throughout the "stable unit treatment value assumption (SUTVA)," which rules out all equilibrium effects; put differently, each unit's treated and untreated outcomes are unaffected by which or how many other units get treated.

In potential outcomes notation, the standard "average treatment effect on the treated" estimand becomes $\text{ATET} = E(Y_1 - Y_0|D = 1) = E(Y_1|D = 1) - E(Y_0|D = 1)$. The treatment group data identify the first term in the

[9] Fraker and Maynard (1987) undertook a similar study using the NSW data around the same time, but they were more focused on comparison group selection than on identification strategies.

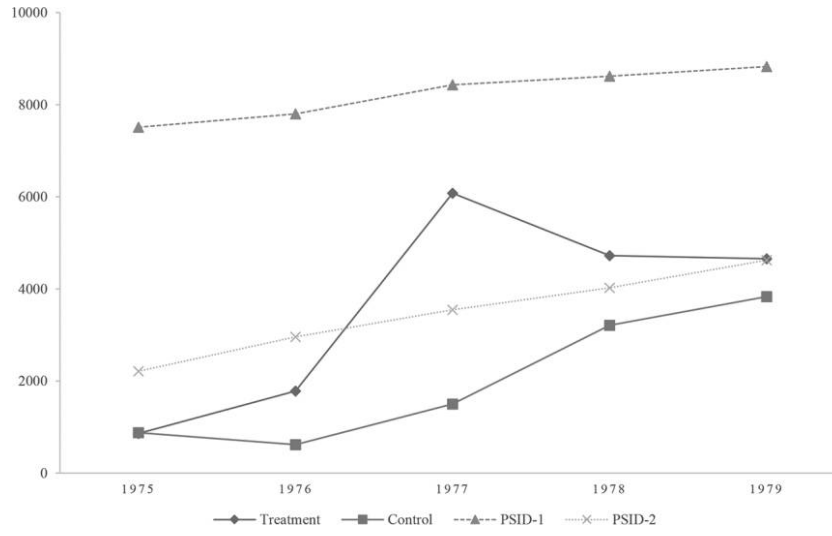

FIG. 1.—Annual earnings: experimental treatment and control groups and PSID-1 and PSID-2 comparison groups. A color version of this figure is available online.

ATET; the second constitutes the always problematic unobserved counterfactual. Experimental evaluations solve the problem of the unobserved counterfactual by forcing would-be treated units (i.e., $D_i = 1$ units) to randomly experience the untreated outcome. Let $R_i \in \{0, 1\}$ indicate random assignment to an experimental treatment group conditional on $D_i = 1$. Then, in an experiment, the population mean outcome for the treatment group $E(Y|D = 1, R = 1)$ corresponds to the first term in the ATET, while the population mean outcome for the control group $E(Y|D = 1, R = 0)$ corresponds to the second term.[10]

While a within-study comparison can examine any partial equilibrium nonexperimental evaluation strategy, to make things concrete, we consider the case of "selection on observed variables." Here the researcher makes a case that for some set of observed covariates for a given comparison group, $E(Y_0|D = 1, X) = E(Y_0|D = 0, X)$. The literature calls this the conditional independence assumption (CIA); it implies that the researcher can condition his/her way out of the problem of nonrandom selection into treatment.[11] Under the CIA, the second term in the ATET corresponds to $\int E(Y_0|D = 0, X = x) f(x|D = 1)dx$.

[10] Given their limited empirical importance in the NSW context, we implicitly assume away treatment group dropout and control group substitution for simplicity.

[11] The usual exogeneity assumption for the parametric linear regression model implies the CIA but not the reverse.

LaLonde (1986) realized that by combining experimental data with a nonexperimental comparison group and treating the experiment as a benchmark, he could examine the performance of particular nonexperimental identification strategies as implemented using specific data sets in the context of a specific program. In particular, he compares experimental impact estimates constructed using the experimental treatment group and the experimental control group to nonexperimental impact estimates constructed by applying a particular identification strategy, such as selection on observed variables, to the experimental treatment group and the nonexperimental comparison group. The difference between these two estimates is the bias associated with the particular nonexperimental estimator in a particular context using a specific comparison group. Heckman et al. (1998) later pointed out that combining the experimental control group and the nonexperimental comparison group provides a second bias estimate.[12]

IV. LaLonde (1986)

LaLonde's (1986) within-study comparison uses the experimental data from the National Supported Work Demonstration described in Section II combined with nonexperimental comparison groups drawn from the Panel Study of Income Dynamics (PSID), a large, nationally representative panel data set, and the Current Population Survey (CPS), a large cross-sectional data set (with a limited panel aspect). LaLonde (1986) combines the men from the dropout, ex-addict, and ex-convict NSW target groups into a single male group. His female group consists solely of the women from the AFDC target group. He then creates comparison groups from the PSID and CPS data sets corresponding to his male and female NSW groups. In this paper, we consider only LaLonde's (1986) AFDC women and (a subset of) the related comparison groups.[13]

[12] For the simplest possible estimator that uses only the mean outcomes in the treatment, control, and comparison groups, the two bias estimates necessarily equal one another because (treatment – control) – (treatment – comparison) = (control – comparison). For more sophisticated estimators, the two bias estimates will differ in finite samples, as indeed they do in our work and in Smith and Todd (2005a, 2005b). To see this, note that, for example, the comparison group nearest neighbors for the treated units may differ from the comparison group nearest neighbors for the control units when applying single nearest neighbor matching. As this discussion suggests, the two estimates are decidedly not independent, but they do provide some information about sensitivity.

[13] Comparison group contamination arises when members of the comparison group receive the treatment under study (or related treatments) but the available data do not note this fact. Such contamination typically arises in contexts like this one wherein standard data sets like the PSID provide the comparison group. Because of the extremely modest size of the NSW Demonstration relative to the population from which we draw PSID-1 and PSID-2, we do not have concerns about contamination in this context.

LaLonde (1986) creates four separate PSID comparison groups for his analysis of the NSW women, of which we consider the two largest. The PSID-1 comparison group includes all female household heads remaining in that status continuously over the period 1975–79 who were between 20 and 55 years old and did not report being retired in 1975.[14] This comparison group represents a random (putting aside issues of nonrandom survey response in the PSID) sample of a much broader population than that implicitly defined by the NSW AFDC women eligibility criteria. LaLonde's three other comparison groups all impose various aspects of the NSW eligibility rules on PSID-1. The PSID unfortunately lacks the covariate detail to impose even an approximate version of the full eligibility criteria.[15]

The PSID-2 comparison group consists of the subset of women in PSID-1 who report receiving any AFDC during calendar year 1975. Absent measurement error in AFDC reporting on the PSID, not a trivial issue empirically, this restriction unambiguously makes the fraction of women in PSID-2 who would have proven eligible for NSW higher than in PSID-1.[16] At the same time, by requiring only that the respondent collect AFDC at some point during 1975, it includes many women with spells too short to meet the NSW eligibility requirement of receiving AFDC in 30 of the previous 36 months.

We replicate LaLonde's (1986) analysis by going back to the raw PSID data and the raw NSW data (both available from the Interuniversity Consortium for Political and Social Research [ICPSR]) and attempting to redo what he did from that point.[17] The appendix, available online, provides a detailed account of our efforts to recreate LaLonde's (1986) analysis files. For the NSW data, we match the sample sizes exactly. The first four columns of table 1 display descriptive statistics for the NSW data; the first two columns repeat the values from table 1 in LaLonde (1986), while the second two present the corresponding values from our data. Similarly, table 2 presents means of the earnings variables from table 2 of LaLonde (1986) and from our data.[18] In general,

[14] This sample restriction represents conditioning on an outcome; it is also not imposed in the NSW data. In fact, around 8% of the NSW women get married by 1979.

[15] Moreover, a PSID comparison sample that met all of the NSW AFDC women eligibility criteria would contain a very modest number of observations, exactly 57 according to footnote 14 of chapter 1 in LaLonde (1984). Presumably, issues such as these motivated the (expensive) collection of a dedicated sample of eligible nonparticipants in the NJS.

[16] On measurement error in AFDC in surveys, including the PSID and several others, see Meyer, Mok, and Sullivan (2015) and the references therein. The measurement error is, of course, highly asymmetric, with many AFDC recipients failing to report receipt and few (if any) nonrecipients reporting receipt.

[17] We leave an analysis of the CPS comparison group to future work.

[18] Nominal earnings were converted to real earnings using the monthly CPI-W reported in the Survey of Current Business. All real earnings throughout the paper are in 1982 dollars, as in LaLonde (1986).

Table 2
Annual Earnings of National Supported Work (NSW) Treatments, Controls, and Panel Study of Income Dynamics (PSID) Comparison Groups

	LaLonde Data				Calónico-Smith Data			
			Comparison Group				Comparison Group	
Year	Treatments	Controls	PSID-1	PSID-2	Treatments	Controls	PSID-1	PSID-2
1975	895	877	7,303	2,327	862	879	7,511	2,211
	(81)	(90)	(317)	(286)	(82)	(91)	(296)	(264)
1976	1,794	646	7,442	2,697	1,783	618	7,801	2,958
	(99)	(63)	(327)	(317)	(95)	(59)	(309)	(318)
1977	6,143	1,518	7,983	3,219	6,077	1,502	8,430	3,545
	(140)	(112)	(335)	(376)	(139)	(111)	(318)	(388)
1978	4,526	2,885	8,146	3,636	4,722	3,212	8,620	4,026
	(270)	(244)	(339)	(421)	(247)	(267)	(328)	(418)
1979	4,670	3,819	8,016	3,569	4,655	3,833	8,827	4,623
	(226)	(208)	(334)	(381)	(227)	(208)	(344)	(513)
No. of observations	600	585	595	173	600	585	648	182

NOTE.—The LaLonde sample was constructed using only participants with valid earnings information in 1975 and 1979. All values are in 1982 dollars. Standard errors are in parentheses.

we match the means on the conditioning variables exactly and come pretty close on the earnings, other than in 1977 and 1978.[19]

For the PSID comparison groups, the story proves a somewhat less happy one, as we end up with noticeably larger PSID samples than in LaLonde (1986). More precisely, our PSID-1 sample has 648 observations compared to 595 in table 2 of LaLonde (1986), and our PSID-2 has 182 compared to his 173, a smaller difference both absolutely and proportionally. The PSID offers multiple measures of household head status, retirement, and some other variables, and neither LaLonde (1986) nor the LaLonde (1984) working paper offer explicit guidance on how he chose among them to the aspiring replicator. As described in the appendix, our earnest explorations along these and other lines did not resolve the mystery. Thus, we proceed with the remainder of our analyses using our best attempt at recreating LaLonde's comparison groups.

LaLonde (1986) does not present descriptive statistics on covariates for the PSID comparison groups, so we can only compare earnings. Table 2 presents the means drawn from table 2 in LaLonde (1986) and those obtained using our samples. Our somewhat larger samples also have somewhat higher mean earnings. For example, in 1979, his PSID-1 has a mean

[19] The sample sizes differ between table 1 and table 2 because we follow LaLonde (1986) in presenting the descriptive statistics for all available observations rather than just for the final analysis sample.

of $8,016, while ours has a mean of $8,827; similarly, for the PSID-2, his sample has a mean of $3,569, and ours has a mean of $4,623. Consistent with earnings having a relatively high variance in these populations (and modest sample sizes), the differences vary by year and sample.

LaLonde (1986) considers three basic identification strategies: selection on observed variables, selection on time-invariant unobserved variables conditional on observed variables, and the bivariate normal selection model, which accounts for selection on time-varying unobserved variables conditional on observed variables under certain (quite strong) parametric assumptions. Following the norm at the time, LaLonde (1986) relies solely on parametric estimators when implementing each identification strategy.

For the selection on observed variables estimator, LaLonde (1986) considers three conditioning sets. The first consists of age, age squared, years of schooling, an indicator for high school dropout status, and indicators for black and Hispanic. The second adds earnings in 1975 to the first. The third controls for "all observed variables," and so apparently adds marital status, residence in a Standard Metropolitan Statistical Area (SMSA), working when surveyed in 1976, and number of children, along with receipt of AFDC in 1975 in some specifications.[20] LaLonde (1986) does not explicitly make a case that these variables suffice for conditional independence, other than noting that the literature as of his writing (e.g., Ashenfelter and Card 1985) emphasizes (as does the more recent literature) the value of conditioning on pre-program outcomes, which suggests an expectation that the first specification will not perform very well.

The second identification strategy assumes "common trends," sometimes termed "bias stability." This identification strategy, when combined with a functional form assumption, motivates application of the parametric linear difference-in-differences estimator. LaLonde (1986) does so both unconditionally and conditional on age to account for nonlinearities in the lifecycle age-earnings profile combined with differences in mean age between the NSW sample and the comparison samples. Implicit in LaLonde's (1986) discussion of the Ashenfelter (1978) dip—the commonly observed pattern that the mean earnings of training program participants decline in the period prior to participation—is the notion of selection on transitory shocks, which in turn suggests that we should not expect very good performance from this estimator in this context; see the detailed discussion of this point in Heckman and Smith (1999).

[20] We say "apparently" because LaLonde (1984, 1986) never makes this covariate set explicit. In our version of this estimator, we do not include SMSA residence because it is not available in the PSID public use data, and we do not include employment as of the 1976 survey because we worry that it constitutes an outcome for some of those randomly assigned very early on.

Finally, LaLonde (1986) applies the Heckman (1979) two-step estimator for the (at the time commonplace) bivariate normal selection model, using various (and no) exclusion restrictions. We do not replicate this approach in our work. First, the two-step estimator is not robust to choice-based sampling, and LaLonde's (1986) combination of NSW and PSID observations represents a decidedly choice-based sample that strongly, but to an unknown extent, overrepresents NSW participants relative to the population.[21] Second, the exclusion restrictions—residence in a SMSA, marital status, employment status in 1976 (after random assignment for some NSW observations), AFDC status in 1975, and number of children—lack face validity, though to be fair, using children as an exclusion restriction was common in female labor supply studies at the time.[22]

The top panel of table 3 presents the estimates from LaLonde's (1986) table 4 based on the NSW women and the PSID-1 and PSID-2 comparison groups, while the bottom panel presents the corresponding estimates using our versions of the NSW AFDC women and the corresponding PSID-1 and PSID-2 comparison groups. We present both experimental and nonexperimental impact estimates, which the reader should compare to one another, and nonexperimental bias estimates, which the reader should compare to zero.[23] The small sample sizes available in the experiment and in the comparison group imply large standard errors for both the experimental and nonexperimental impact estimates and the bias estimates. The experimental impact on earnings in 1979 equals about $851, with a standard error of just over $300, implying a 95% confidence interval of about ($250, $1,450). As such, we can draw only broad conclusions about the performance of alternative nonexperimental strategies; the data require that good performance here roughly means a bias less than or equal to the experimental impact in absolute value.

[21] Footnote 22 in LaLonde (1986) is incorrect in the following sense: the second-stage outcome estimation is robust to choice-based sampling if a population probit underlies the estimation of the selection correction terms. As LaLonde does not weight his probit to undo the choice-based sampling, that is not the case in his application. See the appendix for additional discussion of the choice-based nature of our data.

[22] Because the Wisconsin site, comprising Fond du Lac and Winnebago Counties, served only the AFDC women target group, residence in an SMSA is not a perfect one-way predictor for the AFDC women as it is for LaLonde's sample of men, as pointed out by Smith and Todd (2005a).

[23] As explained in footnote 12, the unconditional bias estimates essentially equal one another: for example, in col. 2, the treatment-control difference in earnings in 1975 equals –$18 and the treatment-comparison difference equals –$6,649 for PSID-1, implying a bias of –$6,631. The control-comparison difference equals –$6,632. The same is true for earnings in 1979 in col. 4. The estimators that involve conditioning yield larger differences in the two alternative bias estimates.

We highlight three major patterns in the estimates. First, our experimental estimates look very similar to those in LaLonde (1986), a not very surprising finding given the close match between his experimental samples and ours documented above. Second, the unadjusted differences in 1975 earnings and 1979 earnings differ only modestly between LaLonde's PSID comparison groups and our versions. For example, the differences for PSID-1 equal –$6,443 in LaLonde (1986) and −$6,649 in column 2 of table 3. In general, but not always, the unadjusted differences get larger rather than smaller in our samples. The same pattern holds for the adjusted (for demographics but not pre-period earnings) differences in columns 3 and 5.

Third, for the difference-in-differences estimators in columns 6 and 7 and the selection-on-observed variables estimators that include pre-program earnings in columns 8–11, we find substantially smaller biases than LaLonde (1986). In column 7, for the PSID-2 comparison group, LaLonde (1986) finds a difference of ($2,392 − $883) = $1,509, compared to ($1,378 − $839) = $539 and $504 using the treatment and control groups with our PSID-2 comparison group. Things look even better with the linear selection-on-observed-variables model in column 9, though they get a bit worse again in columns 10 and 11. In column 9, the biases turn out quite low, less than $250, for both the PSID-1 and PSID-2 compared to both the experimental treatment group and the experimental control group. The reductions in bias starting in column 6 and (more strongly) in column 8 draw attention, yet again, to the value of conditioning flexibly on pre-program outcomes. This strong performance surprises us for (at least) two reasons: for one, our PSID comparison groups do not differ that much in terms of earnings levels from those in LaLonde (1986); for another, these low biases run counter to the claims in Smith and Todd (2005a) regarding the importance of time-invariant differences due to geography and/or earnings measurement between the NSW and the PSID.

V. Dehejia and Wahba (1999, 2002)

The Dehejia and Wahba (1999, 2002) papers innovate in four main ways. First, they focus on a different methodological question, one more about the applied econometrics and less about the economics. While LaLonde (1986) considers the validity of different identification strategies in the NSW context, Dehejia and Wahba assume the validity of a particular identification strategy, namely, conditional independence, and examine the performance of alternative econometric estimators that build on it.

Second, they investigate several matching and weighting estimators based on the estimated probability of participation, or "propensity score," not previously applied to the data from LaLonde (1986) and, indeed, not used very much at all in the empirical economics literature at the time. LaLonde's sample of men represents exactly the context wherein we would expect

Table 3
Earnings Comparisons and Estimated Training Effects for the NSW AFDC Participants Using Comparison Groups from the PSID

		NSW Treatment Earnings Less Comparison Group Earnings									
		Pre-Training, 1975		Post-Training, 1979		Difference-in-Differences in Earnings Growth, 1975–79, Treatments Less Comparisons		Unrestricted Difference-in-Differences: Quasi Difference in Earnings Growth, 1975–79		Controlling for Observed Variables and Pre-training Earnings	
Comparison Group	Comparison Group Earnings Growth 1975–79 (1)	Unadjusted (2)	Adjusted (3)	Unadjusted (4)	Adjusted (5)	Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted (9)	Without AFDC (10)	With AFDC (11)
					LaLonde Data						
Control	2,942	−17	−22	851	861	833	883	843	864	854	· · ·
	(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	1,746	1,354	1,664	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)	· · ·

	Calónico-Smith Data										
Impact estimates:											
Control	2,954	−18	−22	821	841	839	839	824	845	864	···
N = 1,185	(220)	(122)	(122)	(308)	(307)	(324)	(324)	(308)	(306)	(307)	···
PSID-1	1,316	−6,649	−4,894	−4,172	−2,749	2,477	2,261	1,091	664	735	1,013
N = 1,248	(244)	(318)	(327)	(419)	(440)	(342)	(346)	(390)	(410)	(414)	(519)
PSID-2	2,412	−1,350	−1,378	32	−109	1,381	1,378	633	470	478	···
N = 782	(473)	(208)	(215)	(499)	(514)	(504)	(505)	(504)	(519)	(548)	···
Bias estimates:											
PSID-1	1,316	−6,632	−4,957	−4,994	−3,643	1,638	1,351	247	−194	−224	−86
N = 1,233	(244)	(324)	(337)	(413)	(431)	(331)	(334)	(376)	(394)	(397)	(497)
PSID-2	2,412	−1,332	−1,374	−790	−861	542	504	−173	−262	−540	···
N = 767	(473)	(220)	(225)	(470)	(483)	(474)	(474)	(470)	(483)	(504)	···

NOTE.—NSW = National Supported Work; AFDC = Aid to Families with Dependent Children; PSID = Panel Study of Income Dynamics (PSID). Each column presents estimated training effects for a particular econometric model and comparison group. The experimental mean impact estimate is $851 for the LaLonde data and $822 for the Calonico-Smith data. 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 treatment group. The exogenous variables used in the regression-adjusted estimates in cols. 3 and 5 are age, age squared, years of schooling, high school dropout status, marital status, and race. Column 9 adds real earnings in 1975. Finally, cols. 10 and 11 add number of children in 1975. All values are in 1982 dollars. Standard errors are in parentheses.

matching estimators to make a difference relative to parametric linear models with only main effects included. Mechanically, the many incomparable comparison group observations dominate the estimation of the parametric linear model but receive little (if any) weight in the matching and weighting estimators. Third, Dehejia and Wahba highlight the common support (or "overlap") condition and shows its importance in LaLonde's sample of men.

Finally, Dehejia and Wahba consider a subset of LaLonde's sample of men that (arguably) better justifies the conditional independence assumption by allowing them to condition on (more or less) 2 years of pre-random-assignment earnings. In particular, as detailed in Smith and Todd (2005a), they include observations with nonzero earnings in "1974" (really months 13–24 before random assignment) only if randomly assigned in January through April 1976.

Because Dehejia and Wahba did not consider the women in their paper, we cannot know exactly what they would have done with this sample. Imposing the same sample definition Dehejia and Wahba use on the men captures a grand total of only 12 women with nonzero earnings in "1974." Making a compromise between sample size and the spirit of their sample restriction, we expand what we call the "Dehejia-Wahba sample" to include NSW women with nonzero earnings in "1974" randomly assigned anytime in 1976. Table 4, the analogue of table 2 in Smith and Todd (2005a) for the men, illustrates our sample definition.

Table 5 provides descriptive statistics for the Dehejia-Wahba sample of the NSW AFDC women. They reveal a sample quite similar to the original LaLonde sample, the sole substantial difference appearing, by construction, in the month of random assignment. Table 6 shows earnings for the Dehejia-Wahba sample. Compared to the LaLonde sample, both the treatment and control groups have substantially lower earnings in calendar year 1975, around $400, or about half of the value in the larger sample. This difference follows directly from the restrictions imposed in getting from the original LaLonde sample to the Dehejia-Wahba sample. Somewhat surprisingly, this difference largely, but not entirely, disappears in the post-random-assignment period.

We apply two of the matching estimators from Dehejia and Wahba (1999, 2002): propensity score stratification and single nearest neighbor matching with replacement.[24] We pick these particular estimators due to their wide use in the applied literatures in economics and/or statistics. We also apply inverse propensity weighting (IPW), which features in Dehejia and Wahba

[24] Following, for example, Heckman et al. (1998) and much of the applied econometrics literature, we use the term "matching" more broadly than the applied statistics literature, which generally restricts it to what we call single nearest neighbor matching.

Table 4
Sample Composition

Month of Random Assignment	Zero Earnings in Months 13–24 Before RA	Nonzero Earnings in Months 13–24 Before RA	Total
February-76	1	1	2
March-76	12	2	14
April-76	15	3	18
May-76	36	6	42
June-76	31	23	54
July-76	17 **A**	6	23
August-76	33	15	48
September-76	62	15	77
October-76	69	28	97
November-76	55	17	72
December-76	89	28	117
January-77	56	10	66
February-77	93	33	126
March-77	44	17	61
April-77	75	18	93
May-77 **B**	62	**C** 9	71
June-77	57	22	79
July-77	26	11	37
August-77	63	25	88
Total	896	289	1,185

NOTE.—Early random assignment sample = A; $N = 564$. Dehejia-Wahba sample = A+B; $N = 1,040$. LaLonde sample = A+B+C; $N = 1,185$.

Table 5
Descriptive Statistics for Alternative Aid to Families with Dependent Children Experimental Samples

	Dehejia-Wahba Sample		Early Random Assignment Sample	
Variable	Treatments	Controls	Treatments	Controls
Age	34.01	33.91	33.69	35.00
	(7.56)	(7.28)	(7.83)	(7.66)
Years of school	10.25	10.20	10.21	10.11
	(1.89)	(2.10)	(1.93)	(2.11)
Proportion high school dropouts	.71	.69	.72	.71
	(.45)	(.46)	(.45)	(.45)
Proportion married	.02	.04	.02	.04
	(.14)	(.19)	(.16)	(.20)
Proportion black	.83	.80	.87	.86
	(.37)	(.40)	(.33)	(.35)
Proportion Hispanic	.12	.14	.07	.08
	(.32)	(.35)	(.26)	(.26)
Month of assignment (January 1978 = 0)	−12.74	−12.68	−16.04	−16.04
	(4.34)	(4.40)	(2.61)	(2.58)
No. of observations	526	514	285	279

NOTE.—Standard deviations are in parentheses.

Table 6
Annual Earnings of Alternative Aid to Families with Dependent Children Experimental Samples

	Dehejia-Wahba Sample		Early Random Assignment Sample	
Year	Treatments	Controls	Treatments	Controls
1975	377	420	696	773
	(52)	(64)	(93)	(114)
1976	1,703	399	3,093	643
	(101)	(48)	(142)	(83)
1977	6,126	1,486	6,677	1,667
	(150)	(122)	(230)	(188)
1978	4,514	3,184	4,182	2,780
	(264)	(297)	(509)	(434)
1979	4,589	3,800	4,847	3,600
	(242)	(223)	(349)	(307)
No. of observations	526	514	285	279

NOTE.—All values are in 1982 dollars. Standard errors are in parentheses.

(1997), the first chapter of Dehejia's Harvard dissertation. IPW has become a popular choice of applied researchers in recent years. The online appendix provides additional details about these estimators and how we implement them, as well as pointers to the related applied econometric literature.

We use a parametric propensity score model, specifically a logit. We include a flexible specification of the covariates considered in the Dehejia-Wahba papers, namely, age, education (in the form of years of schooling and an indicator for not completing high school), race/ethnicity in the form of indicators for black and Hispanic, marital status, earnings in calendar year 1975 and in "1974," and indicators for zero earnings in 1975 and "1974." We exploit the efficiency gain associated with using both the experimental treatment and control groups in the propensity score estimation throughout our analysis.

Following the literature, for example, Smith and Todd (2005b), Imai, King, and Stuart (2008), and Lee (2013), we undertake a program of balancing tests to choose a specification sufficiently flexible to balance the covariates between the treatment group and the matched (or reweighted) comparison group. The balance tests led us to a model that includes, in addition to main effects in all of the variables that Dehejia and Wahba considered, an interaction between age and years of schooling, which allows for differing age-earnings profiles by years of schooling, as emphasized by Heckman, Lochner, and Todd (2007). The tests also led us to greater flexibility in the conditioning on lagged earnings, where we include squares in earnings from "1974" and 1975, interactions between earnings in "1974" and 1975, and interactions between the indicators for zero earnings in "1974" and 1975. The appendix documents our specification search in greater detail.

Appendix table 2 (appendix tables 1–5 are available online) presents average derivatives from our preferred propensity score models for each comparison group. It contains no substantive surprises.

Figures 2 and 3 display the distributions of the estimated propensity scores using our preferred specification separately for the experimental and comparison group units. Each figure corresponds to one combination of experimental sample (LaLonde, Dehejia-Wahba, or early random assignment) and comparison group (PSID-1 or PSID-2). We note three main patterns. First, for the PSID-1 sample, we see quite substantial separation between the experimental group and comparison group units. Most experimental observations have relatively high estimated propensity scores, while most comparison units have relatively low ones. Second, conditioning the comparison group on AFDC participation in 1975, as we do when going from PSID-1 to PSID-2, makes the distributions of estimated propensity scores dramatically less different. While the experimental units still, as expected, have a distribution of estimated scores with a higher mean and less of a lower tail than the comparison units, the distributions do not differ all that much. Third, even in the case of the PSID-1 comparison group, the common support condition holds more strongly for the AFDC women than for the NSW men, as shown in figures 1 and 2 of Dehejia and Wahba (1999). Both the second and third findings suggest a less difficult selection problem for our data and estimators to solve than that faced in the many papers analyzing the NSW men. See, for example, Crump et al. (2009) for more discussion of common support issues.

Table 7 presents our bias estimates based on the estimators employed by Dehejia and Wahba and obtained using the experimental control group and one of the PSID comparison groups. The reader should compare these bias estimates to zero. The corresponding nonexperimental impact estimates appear in appendix table 4. Within table 7, the first and fourth rows of estimates correspond to the LaLonde NSW sample, the second and fifth to the Dehejia-Wahba sample, and the third and sixth to the early random assignment sample (which we describe in the next section). Similarly, the first three rows of estimates in table 7 correspond to the PSID-1 comparison group and the second three rows to the PSID-2 comparison group. Column 1 in table 7 gives the unconditional mean difference, column 2 the conditional mean difference using a parametric model and the same specification of the covariates used in the propensity scores, column 3 the estimate from propensity score stratification, column 4 the estimate from single nearest neighbor matching with replacement, column 5 the IPW estimate, and column 6 the local linear matching estimate.

All three of the matching estimators substantially reduce the bias relative to the unconditional mean difference for the PSID-1 comparison group. For example, for the LaLonde sample in table 7, the bias falls (in absolute value) from −$4,994 to $609 with single nearest neighbor matching with replacement. Much (much) smaller differences emerge for the PSID-2 sample:

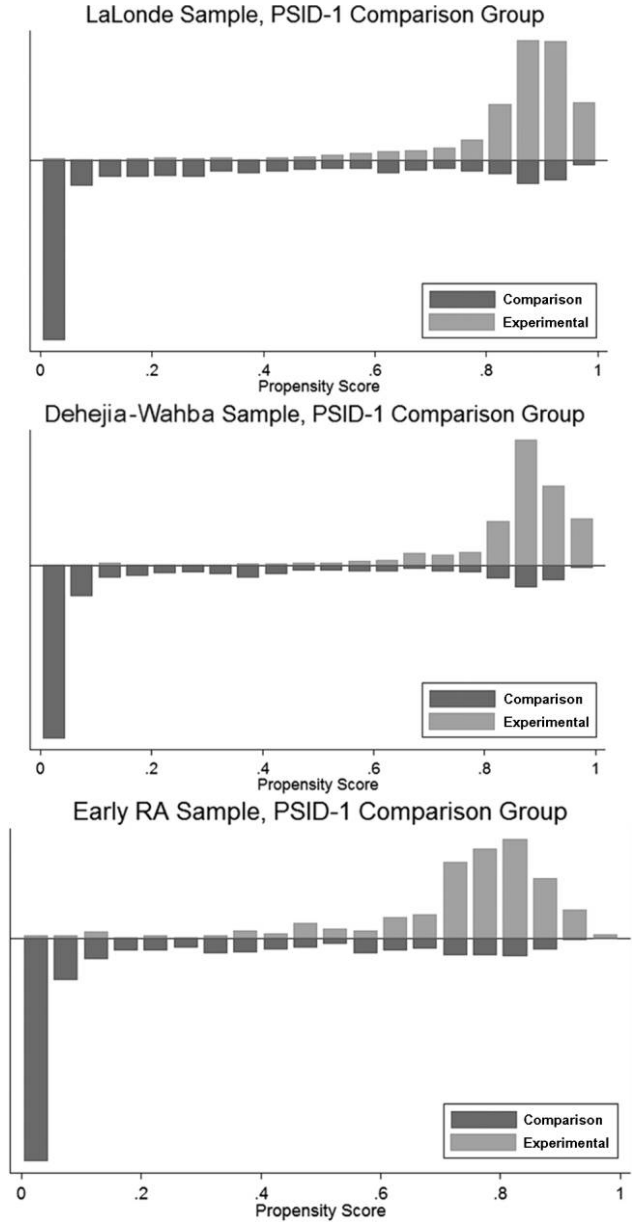

FIG. 2.—Propensity score distributions for the PSID-1 comparison group

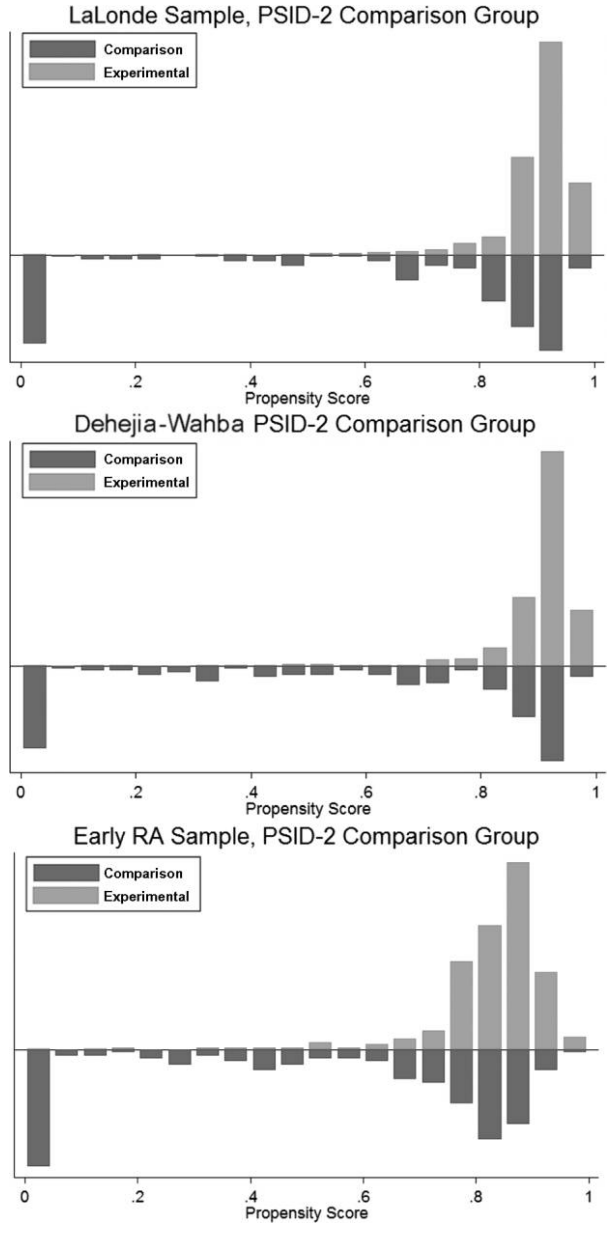

FIG. 3.—Propensity score distributions for the PSID-2 comparison group

Table 7
Bias Estimates Associated with Alternative Cross-Sectional Matching and Weighting Estimators Dependent Variable: Real Earnings in 1979

Sample	Mean Difference (1)	Parametric Linear Model* (2)	Propensity Score Stratification (3)	Single Nearest Neighbor Matching with Replacement (4)	Inverse Propensity Weighting (5)	Local Linear Regression Matching† (6)
	Comparison Group: PSID-1 Female Sample					
LaLonde sample	−4,994	−414	415	609	850	549
	(413)	(423)	(526)	(744)	(562)	(536)
Dehejia-Wahba sample	−5,027	131	757	1,083	1,119	684
	(435)	(462)	(534)	(587)	(556)	(573)
Early random assignment sample	−5,227	−72	280	375	449	109
	(562)	(503)	(547)	(585)	(534)	(616)
	Comparison Group: PSID-2 Female Sample					
LaLonde sample	−790	−419	400	279	641	450
	(470)	(524)	(782)	(1,008)	(663)	(695)
Dehejia-Wahba sample	−823	−98	705	517	1,000	796
	(484)	(555)	(647)	(719)	(604)	(621)
Early random assignment sample	−1,023	−367	230	−249	131	92
	(562)	(602)	(722)	(907)	(703)	(671)

NOTE.—Under the conditional independence assumption, the population value of the bias equals zero. All values are in 1982 dollars. Standard errors are in parentheses.
* The least squares regression in col. 2 uses the same specification as the propensity score model from appendix table 2.
† LLR matching uses ROT bandwidth selector from Stata command lpoly. The computed values for each row are .176, .166, .123, .178, .117, and .167.

again for the LaLonde sample, the bias falls (in absolute value) from −$790 to $279 for the same estimator. The biases turn out to be similar in magnitude to the experimental impact estimates; this makes them substantively moderate but still a bit too large for one to want to rely on nonexperimental evaluation in this context. They also have, as in Dehejia and Wahba (1999, 2002) and Smith and Todd (2005a, 2005b), relatively large standard errors, reflecting the smallish samples and the large residual variance of earnings in this population.

Compared to the parametric linear regression estimators in column 2 of table 7, which contains the exact same specification of the conditioning variables as in the propensity score model, the matching estimators typically yield somewhat larger bias estimates. For example, for the Dehejia-Wahba sample, the rich selection-on-observed variables specification in column 2 yields a bias of $131 with the PSID-1 and −$98 with the PSID-2, in contrast to biases for the single nearest neighbor estimate with propensity scores based on the same covariate set of $1,083 and $517 for the PSID-1 and PSID-2 comparison groups, respectively.[25] As expected, the parametric estimators also yield smaller standard errors. The relative performance of the matching estimates to the parametric linear regression estimates with the same covariate sets clearly differs from that found for the men in Dehejia and Wahba (1999). We attribute this change in performance to the less difficult selection problem posed by the AFDC women, as illustrated by the much more similar distributions of estimated propensity scores, particularly for the PSID-2 comparison group. Put differently, the women lack the strongly asymmetric distributions of propensity scores and huge mass of comparison observations with propensity scores near zero that characterize the men, and thus they do not fit the profile of a case where matching should make a big difference.

The IPW estimates look quite similar to the matching estimates in magnitude. At the same time, IPW rather clearly decreases the estimated standard errors by making more efficient use of the available data. For those concerned with mean squared error rather than just bias, this pattern reaffirms the general finding from the Monte Carlo literature that IPW dominates single nearest neighbor matching.

Turning to secondary findings, as with the parametric estimates in table 3, we find noticeably lower bias estimates when using the PSID-2 comparison group than when using the broader PSID-1 comparison group. This provides additional support to the view that imposing even partial eligibility criteria on the comparison group reduces the severity of the selection problem that the econometric estimators have to deal with. Finally, we find larger

[25] Angrist (1998) notes a fact still not widely recognized in the applied literature: matching and parametric linear regression have different causal estimands. We do not expect that difference in estimands to account for much of the difference in bias estimates we describe.

biases in general for the Dehejia and Wahba–inspired sample than for the LaLonde sample, which differs strongly from the pattern found for the men by Smith and Todd (2005a). For them, the Dehejia-Wahba sample exclusion lessens the selection problem by moderating the pre-program earnings dip; the AFDC women have no real dip to moderate due to their ongoing welfare participation. We conclude with a final reminder that, as always with the NSW data, we suffer terribly from large standard errors; the patterns described in this section show up clearly in the point estimates but likely often do not achieve statistical significance at standard levels.

VI. Smith and Todd (2005a, 2005b)

Smith and Todd (2005a, 2005b) stand on the shoulders of LaLonde (1986) and Dehejia and Wahba (1999, 2002) and add to this literature in several ways. First, they consider a second subsample of the LaLonde data that, like the Dehejia-Wahba sample, allows for conditioning on 2 years of pre-random-assignment earnings but, unlike the Dehejia-Wahba sample, does not treat observations asymmetrically based on whether their earnings in months 13–24 before random assignment equal zero or not. Second, Smith and Todd apply the difference-in-differences matching methods developed in Heckman et al. (1998). Third, they apply the kernel and local linear matching estimators developed in Heckman, Ichimura, and Todd (1997, 1998). These estimators have important advantages relative to the propensity score stratification and nearest neighbor matching estimators applied in Dehejia and Wahba (1999, 2002). Fourth, they consider "pre-program" tests of overidentifying restrictions like those examined in Heckman and Hotz (1989). Fifth, they develop and apply an alternative balance test procedure and show that some of the Dehejia-Wahba specifications that pass their balance test fail the Smith-Todd balance test. Finally, they examine the sensitivity of the Dehejia-Wahba estimates to a variety of minor implementation changes, such as the handling of ties in the nearest neighbor matching and whether the propensity score estimation relies on the experimental treatment group, the experimental control group, or both.

Of these six contributions, we focus in this section on just three: the early random assignment sample, the alternative matching estimators, and the difference-in-differences matching. We incorporate the various lessons learned from the sensitivity analyses throughout the paper. We do not consider the Smith-Todd (2005b) balancing test for reasons outlined in the appendix. We do not consider the pre-program tests because we view them as less informative than Smith and Todd (2005a) do.[26]

[26] In particular, because the propensity score specifications adopted in Smith and Todd (2005a, 2005b) include earnings in 1975, a pre-program test using earnings in 1975 represents a balance test, rather than a test of identifying assumptions. A pre-program test aimed at identification would use propensity scores that included earnings variables lagged relative to 1975. Sadly, the NSW data lack such variables.

The Smith-Todd (2005a) early random assignment sample for the NSW men includes individuals randomly assigned in January through April (inclusive) of 1976, that is, during the first 4 months of random assignment. This sample includes just 108 treated units and 142 control units, 78 and 118 fewer than the Dehejia-Wahba male samples, respectively. The benefit that Smith and Todd (2005a) think offsets this cost in sample size comes from not having to treat individuals asymmetrically based on their earnings in months 13–24 before random assignment. For all of the observations in their early random assignment sample, earnings in "1974" come pretty close to earnings in 1974. Thus, they address Dehejia and Wahba's (1999, 2002) valid concern about having 2 years of pre-random-assignment earnings to condition on, and they do so (in some sense) even more strongly than Dehejia and Wahba do. The early random assignment sample also improves the temporal alignment between the time-varying conditioning variables and the choice to participate in the NSW.

In the context of the NSW women, we face a nasty trade-off. Random assignment got going more slowly for the women in the NSW experiment than it did for the men. If we restrict ourselves to women randomly assigned in just the first 4 months of 1976, we have only 66 treatment group observations and 64 control observations, which are small numbers indeed even by the low standards of the NSW literature. Thus, relative to Smith and Todd (2005a), we trade some match quality between earnings in "1974" and 1974 and some temporal alignment for some additional sample size by defining our early random assignment sample as including all women randomly assigned anytime in 1976, which includes 285 treatment group members and 279 controls.

Table 5 provides descriptive statistics on the early random assignment sample, and table 6 describes their earnings for calendar years 1975–79. Relative to the LaLonde and Dehejia-Wahba samples, their characteristics differ very little, other than having a slightly higher proportion black and a slightly lower proportion Hispanic. In terms of earnings, the early random assignment sample looks more like the LaLonde sample than the Dehejia-Wahba sample, a not surprising finding given serially correlated earnings and the fact that the Dehejia-Wahba sample omits individuals with nonzero earnings in months 13–24 prior to random assignment. For unknown reasons, the early random assignment experimental impact well exceeds that of either of the other two experimental samples.

Following Heckman et al. (1998) and Heckman, Ichimura, and Todd (1997, 1998), Smith and Todd (2005a, 2005b) consider matching estimators based on kernel or local linear regressions (LLR) of the untreated outcomes of the comparison group units on the estimated propensity scores. We limit ourselves to the LLR matching estimator for technical reasons, which are outlined in the appendix. Column 6 of table 7 presents the bias estimates, and appendix table 4 presents the analogous nonexperimental impact estimates. The LLR estimator reduces the bias relative to the nearest neighbor

estimators for four out of six combinations of NSW sample and comparison group, but modestly so. It yields standard errors similar in size to those provided by IPW.

Finally, we consistently find lower (in absolute value) biases in the estimates for the early random assignment sample than for either the Dehejia-Wahba or the LaLonde samples. This pattern diverges sharply from the parallel analysis in Smith and Todd (2005a), which consistently found much larger biases for the early random assignment sample. There it seems that the early random assignment sample posed a more challenging selection problem due to the lower representation of individuals with zero earnings throughout the pre-random-assignment period among the NSW men. Large falls in earnings during the period prior to participation represent (much) less of an issue for the NSW AFDC population. Instead, we conjecture that the estimates reflect bias reductions resulting from better temporal alignment of the pre-program earnings variables and the timing of random assignment (and, thus, the choice to participate in NSW).

Difference-in-differences matching extends the traditional linear parametric difference-in-differences estimator to allow for semi-parametric conditioning on the covariates. It builds on an assumption of conditional bias stability—that, conditional on some set of exogenous covariates, the difference in mean untreated outcomes between the treated and untreated units remains constant over time, at least over the period covered by the analysis. Put differently, difference-in-differences deals with nonrandom selection into treatment that depends on both observed covariates and time-invariant unobserved variables.

The recent literature includes some debate about whether to prefer differences-in-differences to flexible conditioning on pre-program outcomes. Substantively, a sufficiently rich set of pre-program outcomes should capture both time-varying and time-invariant unobserved factors affecting participation and outcomes. This seems to be the case in, for example, Andersson et al. (2013), who obtain roughly the same estimates with both approaches in their nonexperimental evaluation of Workforce Investment Act training. Imbens and Wooldridge (2009) and Chabé-Ferret (2015) provide further discussion.

Table 8 and appendix table 5 display the estimates based on difference-in-differences matching. The estimators parallel those in table 7 with the omission of the parametric linear model, which provides the same estimates in both cases due to the inclusion of earnings in 1975 in the conditioning set. In all cases, calendar year 1979 earnings serve as the "post" period, and calendar year 1975 earnings serve as the "pre" period. Regarding the estimators, the same patterns emerge that we saw in table 7 with the cross-sectional estimators: not much of a systematic effect of IPW or LLR on the estimated bias but smaller standard errors for IPW and LLR than for the nearest neighbor estimators. Similarly, we obtain the same pattern across samples, with

Table 8
Bias Estimates Associated with Alternative Difference-in-Differences Matching and Weighting Estimators: Dependent Variable = Difference between Real Earnings in 1979 and Real Earnings in 1975

Sample	Mean Difference (1)	Propensity Score Stratification* (2)	Single Nearest Neighbor Matching with Replacement (3)	Inverse Propensity Weighting (4)	Local Linear Regression Matching† (5)
	Comparison Group: PSID-1 Female Sample				
LaLonde sample	1,638	363	430	660	432
	(331)	(539)	(722)	(555)	(523)
Dehejia-Wahba sample	2,064	786	977	1,044	834
	(342)	(537)	(589)	(534)	(538)
Early random assignment sample	1,511	217	419	312	289
	(427)	(567)	(566)	(517)	(538)
	Comparison Group: PSID-2 Female Sample				
LaLonde sample	542	118	124	476	13
	(474)	(734)	(1,002)	(586)	(658)
Dehejia-Wahba sample	969	619	449	905	639
	(478)	(651)	(708)	(608)	(640)
Early random assignment sample	415	−71	−436	−16	−133
	(550)	(715)	(916)	(693)	(708)

NOTE.—Under the conditional independence assumption, the population value of the bias equals zero. All values are in 1982 dollars. Standard errors are in parentheses.
* The least squares regression in col. 2 uses the same specification as the propensity score model from appendix table 2.
† LLR matching uses ROT bandwidth selector from Stata command lpoly. The computed values for each row are .192, .201, .242, .160, .213, and .157.

smaller estimated biases for the early random assignment sample relative to the LaLonde and Dehejia-Wahba samples.

Juxtaposing the cross-sectional matching estimates in table 7 and the difference-in-differences matching estimates in table 8 reveals that the latter typically have a modestly lower bias than the former. This suggests a relatively minor substantive gain to removing time-invariant differences due to geographic mismatch and/or systematic differences in earnings measurement between the NSW data and the PSID. Smith and Todd (2005a) emphasize these factors in explaining the much larger, in both absolute and relative terms, bias reduction associated with difference-in-differences matching relative to cross-sectional matching for the men in their study. While we might expect these factors to matter somewhat less for the NSW AFDC women due to their lower earnings levels, the contrast exceeded our expectation and calls into question the emphasis that Smith and Todd (2005a) put on these explanations.

VII. Conclusion

Our replication of LaLonde's (1986) analysis of the NSW AFDC women, and extension of the related analyses in Dehejia and Wahba (1999, 2002) and Smith and Todd (2005a, 2005b) to the NSW AFDC women, provides a number of valuable lessons. In terms of replicating the original LaLonde (1986) paper, we had no trouble with the NSW experimental data, but we did not succeed in exactly replicating LaLonde's PSID comparison group samples. We lack the information, such as LaLonde's original data-cleaning programs and the version of the PSID that he used, which would be required to pin down the exact source of the differences. Our troubles highlight the value of the recent practice at many journals of requiring authors to deposit their code when their paper gets published.

Conceptually, the paper has provided an opportunity to illustrate alternative notions of research replication and extension. It has also allowed us to make the important distinction, sometimes missed in the literature, that undertakes within-study comparisons between learning about the plausibility of particular identifying assumptions in the context of particular institutions and data sets and learning about the performance of particular estimators that rely on the same basic identifying assumption. The former represents a substantive economic question, the latter an applied econometric or statistical question.

In regard to the applied econometrics, we find, like Smith and Todd (2005a), that alternative matching and weighting estimators do not deliver large differences in estimated biases. However, the IPW estimator does pay off in terms of large reductions in variance. This finding comports with the picture painted by the Monte Carlo literature, such as Huber, Lechner, and Wunsch (2013) and Busso, DiNardo, and McCrary (2014).

Finally, in terms of the substance, we offer five major conclusions. First, our version of LaLonde's analysis file yields the same qualitative conclusions (and often quite similar quantitative conclusions) as he obtained in his original paper.

Second, the AFDC women pose an easier selection problem than the NSW men, because they have a less dramatic pre-program dip in earnings and are generally more homogeneous. This shows up in the common support graphs for the estimated propensity scores, particularly in the case of the PSID-2 comparison group. These graphs show much less separation than the corresponding graphs for the NSW men in Dehejia and Wahba (1999, 2002) and Smith and Todd (2005a). It also shows up in the fact that going from parametric linear regression models to propensity score stratification and matching estimators has little effect for the NSW AFDC women, again unlike the men. Matching has the most potential to matter to the estimates when the comparison group contains many incomparable observations. This condition holds for the NSW men but not for the NSW women.

Third, the low bias estimates associated with the PSID-2 comparison group also support imposing, to the extent feasible, the program eligibility rules when constructing comparison groups for evaluations of voluntary programs that rely on identification via conditional independence.

Fourth, the very different pattern of biases that we find among the LaLonde, Dehehia-Wahba, and early random assignment samples than Smith and Todd (2005a) find for the men suggests the importance of issues of temporal alignment between time-varying conditioning variables and the choice to participate in the program or not. Dolton and Smith (2011) emphasize this issue in their context, as does the recent literature on dynamic treatment effects initiated by Sianesi (2004). Temporal alignment suggests lower bias for the early random assignment sample as earnings in 1975 occur reasonably close to the participation choice for them, but not for the observations randomly assigned in 1977 included in both the Dehejia-Wahba and LaLonde samples. We suspect that temporal alignment matters for the men as well but that its effects get swamped by the much harder selection problem they exhibit.

Fifth, and finally, we do not find large differences in bias estimates between the estimators based on the conditional independence assumption, for example, the matching and IPW estimators, and the estimators based on the conditional bias stability assumption, for example, difference-in-differences matching and IPW using the before-after earnings difference as the dependent variable. Smith and Todd (2005a) found large differences between the estimates built on these two identification strategies, with much lower bias for the difference-in-differences estimators, and they interpreted them as signaling the importance of time-invariant differences between the NSW and comparison group observations resulting from differences in the measurement of earnings in the NSW data and the PSID and/or differences in

earnings resulting from geographic mismatch between the NSW sites and the PSID. Our findings for the NSW AFDC women suggest that these represent only minor factors and that future research should look for an explanation specific to the men, rather than one that applies to both men and women.

What does our analysis have to say about the larger question of how policy makers should think about the plausibility of nonexperimental estimates? We view the underlying research problem as learning what identification strategies and estimators produce reasonable estimates for particular combinations of program type (e.g., employment and training programs) and data (e.g., many years of pre-program earnings or not). The literature on within-study designs, to which this paper contributes, represents one very important way in which to accumulate knowledge relevant to this research question. Indeed, several of the substantive conclusions just reviewed have real relevance to policy makers (or, more realistically, their research-consuming policy wonk assistants) seeking to weigh the evidence on a given program (or to fund an evaluation going forward). They should put more weight on studies that use comparison groups eligible for the program under study, should worry less about selection when studying voluntary employment and training programs for women on welfare than when studying similar programs for men, and should assign more weight to studies that pay attention to the temporal alignment between time-varying conditioning variables and the program participation choice. More deeply, they should emphasize studies that make an explicit case for their identification strategy, a case that, in many contexts, can build on a stock of knowledge produced by within-study designs.

We end with a suggestion for improving the within-study design literature. As Michael Lechner frequently reminds (one of) us, the NSW data have really small sample sizes and, thus, really large standard errors, especially when combined, as in this paper, with the smaller of the two comparison groups considered by LaLonde (1986), namely, that from the PSID. Plus, arguably, conditional independence holds (roughly) for the NSW women but not for the NSW men. We think these facts imply that the methodological literature in applied econometrics (and more recently in the causal part of applied statistics) really should find an alternative, larger, canonical data set for examining the performance of nonexperimental identification strategies and estimators.

References

Andersson, Fredrik, Harry Holzer, Julia Lane, David Rosenblum and Jeffrey Smith. 2013. Does federally-funded job training work? Nonexperimental estimates of WIA training impacts using longitudinal data on workers and firms. NBER Working Paper no. 19446, National Bureau of Economic Research, Cambridge, MA.

Angrist, Joshua. 1998. Estimating the labor market impact on voluntary military service using social security data on military applicants. *Econometrica* 66, no. 2:249–88.

Ashenfelter, Orley. 1978. Estimating the effect of training programs on earnings. *Review of Economics and Statistics* 60, no. 1:47–57.

Ashenfelter, Orley, and David Card. 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics* 67, no. 4:648–60.

Busso, Matias, John DiNardo, and Justin McCrary. 2014. New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics* 96, no. 5:885–97.

Chabé-Ferret, Sylvain. 2015. Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes. *Journal of Econometrics* 185, no. 1:110–23.

Clemens, Michael. 2015. The meaning of failed replications: A review and proposal. IZA Discussion Paper no. 9000, Institute of Labor Economics, Bonn, Germany.

Couch, Kenneth. 1992. New evidence on the long-term effects of employment and training programs. *Journal of Labor Economics* 10, no. 4:380–88.

Crump, Richard, V. Joseph Hotz, Guido Imbens, and Oscar Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96, no. 1:187–99.

Dehejia, Rajeev, and Sadek Wahba. 1997. Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. In Econometric methods for program evaluation, PhD diss., Harvard University.

———. 1999. Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, no. 448:1053–62.

———. 2002. Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84, no. 1:151–61.

Dolton, Peter, and Jeffrey Smith. 2011. The impact of the UK New Deal for Lone Parents on benefit receipt. IZA Discussion Paper no. 5491, Institute of Labor Economics, Bonn, Germany.

Duvendack, Maren, Richard Palmer-Jones, and W. Robert Reed. 2015. Replication in economics: A progress report. *Econ Journal Watch* 12, no. 2: 161–94.

Fraker, Thomas, and Rebecca Maynard. 1987. The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources* 22, no. 2:194–227.

Friedlander, Daniel, and Philip Robins. 1995. Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review* 85, no. 4:923–37.

Hamermesh, Daniel. 2007. Viewpoint: Replication in economics. *Canadian Journal of Economics* 40, no. 3:715–33.

———. 2016. Replication in labor economics: Evidence from data, and what it suggests. IZA Discussion Paper no. 10403, Institute of Labor Economics, Bonn, Germany.

Heckman, James. 1979. Sample selection bias as a specification error. *Econometrica* 47, no. 1:153–61.

Heckman, James, Neil Hohmann, and Jeffrey Smith, with the assistance of Michael Khoo. 2000. Substitution and drop out bias in social experiments: A study of an influential social experiment. *Quarterly Journal of Economics* 115, no. 2:651–94.

Heckman, James, and V. Joseph Hotz. 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association* 84, no. 408:862–80.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. Characterizing selection bias using experimental data. *Econometrica* 66, no. 5:1017–98. [Note: All references in the text to Heckman et al. (1998) are to this item.]

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies* 64, no. 4:605–54.

———. 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65, no. 2:261–94.

Heckman, James, Robert LaLonde, and Jeffrey Smith. 1999. The economics and econometrics of active labor market programs. In *Handbook of labor economics*, vol. 3A, ed. Orley Ashenfelter and David Card, 1865–2097. Amsterdam: Elsevier.

Heckman, James, Lance Lochner, and Petra Todd. 2007. Earnings functions, rates of return, and treatment effects: The Mincer equation and beyond. In *Handbook of the economics of education*, vol. 1, ed. Eric Hanushek and Finnis Welch, 307–458. Amsterdam: Elsevier.

Heckman, James, and Jeffrey Smith. 1999. The pre-program earnings dip and the determinants of participation in a social program: Implications for simple program evaluation strategies. *Economic Journal* 109, no. 457: 313–48.

Hollister, Robinson. 1984. The design and implementation of the Supported Work evaluation." In Hollister et al. 1984, 12–49.

Hollister, Robinson, Peter Kemper, and Rebecca Maynard. 1984. *The National Supported Work Demonstration*. Madison: University of Wisconsin Press.

Hollister, Robinson, and Rebecca Maynard. 1984. The impacts of Supported Work on AFDC recipients. In Hollister et al. 1984, 90–135.

Hollister, Robinson, and Elizabeth Wilde. 2007. How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management* 26, no. 3: 455–77.

Huber, Martin, Michael Lechner, and Conny Wunsch. 2013. The performance of estimators based on the propensity score. *Journal of Econometrics* 175, no. 1:1–21.

Imai, Kosuke, Gary King, and Elizabeth Stuart. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 171 (pt. 2):481–502.

Imbens, Guido, and Jeffrey Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47, no. 1:5–86.

Kemper, Peter, David Long, and Craig Thornton. 1981. *The Supported Work Evaluation: Final cost benefit analysis*. New York: Manpower Demonstration Research Corporation.

LaLonde, Robert. 1984. Evaluating the econometric evaluations of training programs with experimental data. Industrial Relations Section Working Paper no. 183, Princeton University.

———. 1986. Evaluating the econometric evaluations of training programs using experimental data. *American Economic Review* 76, no.4:604–20.

Lee, Wang-Sheng. 2013. Propensity score matching and variations on the balancing test. *Empirical Economics* 44, no. 1:47–80.

Lise, Jeremy, Shannon Seitz, and Jeffrey Smith. 2004. Equilibrium policy experiments and the evaluation of social programs. NBER Working Paper no. 10283, National Bureau of Economics Research, Cambridge, MA.

Meyer, Bruce, Wallace Mok, and James Sullivan. 2015. The under-reporting of transfers in household surveys: Its nature and consequences. Unpublished manuscript, University of Chicago.

Sianesi, Barbara. 2004. An evaluation of the Swedish system of active labor market programs in the 1990s. *Review of Economics and Statistics* 86, no. 1: 133–55.

———. 2017. Evidence of randomisation bias in a large-scale social experiment: The case of ERA. *Journal of Econometrics* 198, no. 1:61–64.

Smith, Jeffrey, and Petra Todd. 2005a. Does matching overcome LaLonde's critique of nonexperimental methods? *Journal of Econometrics* 125, nos. 1–2:305–53.

———. 2005b. Rejoinder. *Journal of Econometrics* 125, nos. 1–2:365–75.

Todd, Petra, and Kenneth Wolpin. 2006. Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *American Economic Review* 96, no. 5:1384–1417.