The Women of the National Supported Work Demonstration

Sebastian Calónico
University of Miami
scalonico@gmail.com

Jeffrey Smith
University of Michigan, NBER and IZA
econjeff@umich.edu

August 14, 2015

We thank participants at the conference in honor of Robert LaLonde held at the Chicago Fed in April 2015 for helpful comments, particularly our discussant Jens Ludwig. We also thank participants in the 2009 Danish Microeconometric Network meetings in Copenhagen in 2009, the 3rd Joint IZA/IFAU Conference on Labor Market Policy Evaluation in Uppsala in 2010 and at the "Q" Center at Northwestern in 2013 for valuable feedback. Louis-Pierre Lepage provided valuable research assistance with enthusiasm and interest. The reasoning in this paper has benefitted from many general discussions over the years with James Heckman, Dan Black, Michael Lechner, Petra Todd, Barbara Sianesi, and Jessica Goldberg. Much as we might like to blame any errors on anonymous passers-by, we must instead retain responsibility for any that might appear. Please tell us about them.

1. Introduction

Within-study comparisons using experiments as benchmarks against which to judge the performance of non-experimental identification strategies applied using particular datasets 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 Support Work Demonstration with non-experimental 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), studies that we describe in more detail below, the within-study comparison literature includes a number of papers based on the data from the National Job Training 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; examples include Heckman and Smith (1999) and Heckman, Ichimura, Smith, and Todd [hereinafter HIST] (1998). 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 non-experimental identification strategies work with what data in particular institutional contexts, as well as about the performance of structural models (in the sense that economists use that term).

This study replicates LaLonde's (1986) analysis of the Aid to Families with Dependent Children (AFDC) women target group.[1] Because these data were lost, but the data on LaLonde's male sample were 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 also repeat the analyses in Dehejia and Wahba [hereinafter DW] (1999, 2002) and Smith and Todd [hereinafter ST] (2005a,b) using our reconstruction of LaLonde's data on women. The DW (1999, 2002) papers played a crucial role in introducing matching estimators from the applied statistics literature into the applied economics literature. ST (2005a,b) helped to popularize the difference-in-differences matching estimators introduced in HIST (1998) and, more importantly, curbed the enthusiasm of the empirical literature for propensity score matching engendered by overly-optimistic readings of DW (1999, 2002) by showing the sensitivity of their cheery conclusions to aspects of both the sample and the estimation. Repeating the analyses in these papers 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 DW (1999, 2002) and ST (2005a,b) to an additional target group. The conceptual "big tent" of replication includes many types of analyses, ranging from simply re-running 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, more prosaically, neither LaLonde's (1986) code nor his analysis file for the AFDC women have survived to the present.

[1] AFDC is the predecessor of the current Temporary Assistance for Needy Families (TANF) program.

Finally, we address two additional issues related to the substance of active labor market program evaluation. The literature in the US contains hints that AFDC women represent a less challenging evaluation problem than do disadvantaged men; see e.g. the findings in Friedlander and Robins (1995). They typically have a less dramatic pre-program dip in earnings 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. We consider the contrast between our findings and those of the papers examining the men in this light. In addition, by contrasting our findings for comparison groups that do and do not impose (some of) the eligibility requirements for the NSW treatment, we shed additional light on the value of restricting comparison groups solely to program eligibles.

The remainder of the paper leads the reader down the following path: Section 2 describes the now ancient National Supported Work Demonstration, which provides our experimental data. Section 3 details the conceptual framework for within-study comparisons. Section 4 describes the LaLonde (1986) study in general and, in more detail, our replication of his analysis of the long-term AFDC women target group. Section 5 presents our extension of DW (1999, 2002) and Section 6 does the same for our extension of ST (2005a, b). Section 7 concludes.

2. 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 fifteen locations throughout the United States. Four target groups were selected for inclusion in the program: female long-term AFDC recipients, former drug addicts, unemployed ex-offenders, and young school dropouts. The program first 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 $14000 in direct program operating costs in 1997 dollars – than typically spent under other programs such as the Workforce Investment Act (WIA).[2]

To participate in NSW, a set of eligibility criteria was established, in order 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 four weeks preceding the time of selection into the program), and (2) the person must have spent no more than three months on one regular job of at least 20 hours per week during the preceding six months. For the AFDC target group, additional criteria applied: (3) no child age less than six years; and (4) on AFDC for at least 30 of the last 36 months.

From April 1975 to August 1977, the NSW demonstration operated as a randomized experiment in 10 of its 15 cities, including eight sites serving AFDC women. Along with the Negative Income Tax (NIT) experiments, the NSW represented one of the first major social experiments in the US (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 four follow-up surveys.[3] Couch (1992) provides long-term impact estimates for LaLonde's male and female samples using administrative data. He finds persistent positive impacts for the AFDC women and persistent zeros for the men. 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.

[2] See e.g. the discussions around Table 18 of Heckman, LaLonde and Smith (1999) and the references therein.
[3] These interviews cover the two years prior to random assignment and every nine months thereafter (up to 36 months, or four post-baseline interviews). The data provide information on demographic characteristics, employment history, job search, mobility, household income, housing and drug use. The NSW administrators also scheduled a 27th-month interview for only 65 percent of the participants and a 36th-month interview for 24 percent of the non-AFDC participants. None of the AFDC participants were scheduled for a 36th-month interview, but instead, a resurvey during 1979 including 75 percent of these women anywhere from 27 to 44 months after the baseline. Response rates were an issue; see the discussion in LaLonde (1986).

3. Within-study comparisons

The deepest contribution of LaLonde (1986) consists of his introduction of what the literature has come to call "within-study" comparisons.[4] Such comparisons use experimental evaluations as benchmarks for the performance of non-experimental estimators of various sorts applied to data on non-experimental 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 person "*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 $Y_i = D_i Y_{1i} + (1 - D_i)Y_{0i}$. In words, observational data provide the treated outcome for the treated units and the untreated outcome for the untreated units. 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 $ATET = E(Y_1 - Y_0 \mid D = 1) = \mathrm{E}(\mathrm{Y}_1 \mid D = 1) - E(Y_0 \mid D = 1)$. The treatment group data identify the first term in the 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

[4] Fraker and Maynard (1987) undertook a similar study using the NSW data around the same time, but more focused on comparison group selection than on identification strategies. We thank Tom Cook for a helpful email exchange on the intellectual history of within-study comparisons.

conditional on $D_i = 1$. Then, in an experiment the population mean outcome for the treatment group $E(Y \mid D = 1, R = 1)$ corresponds to the first term in the ATET while the population mean outcome for the control group $E(Y \mid D = 1, R = 0)$ corresponds to the second term.[5]

While a within-study comparison can examine any partial equilibrium non-experimental 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, $E(Y_0 \mid D = 1, X) = E(Y_0 \mid D = 0, X)$. The literature calls this the conditional independence assumption (CIA) or, in the awkward terminology of applied statistics, unconfoundedness.[6] Under the CIA, the second term in the ATET corresponds to $\int E(Y_0 \mid D = 0, X = x) f(x \mid D = 1) dx$. In words, under the CIA the researcher can condition her way out of the problem of non-random selection into treatment.

LaLonde (1986) realized that by combining experimental data with a non-experimental comparison group and treating the experiment as a benchmark, he could examine the performance of particular non-experimental 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 non-experimental estimates constructed using the experimental treatment group and the non-experimental comparison group, such as the selection on observed variables strategy just described. The difference between the two represents an estimate of the bias associated with the particular non-experimental estimator as applied in a particular context using particular data. HIST (1998) later pointed out that combining the experimental control group and the non-

[5] We implicitly assume away treatment group dropout and control group substitution for simplicity. See the extended discussion in Heckman, Hohmann, Smith and Khoo (2000).
[6] The usual exogeneity assumption for the parametric linear regression model implies the CIA, but not the reverse.

experimental comparison group provides a second bias estimate. The "within-study" nomenclature reflects the fact that the experimental and non-experimental estimates share either the experimental treatment group or the experimental control group; in that partial but important sense, the two come from within the same study.

4. Replicating LaLonde (1986)

LaLonde's (1986) within-study comparison uses the experimental data from the National Supported Work (NSW) demonstration described in Section 2 combined with non-experimental comparison groups drawn from the Panel Study of Income Dynamics (PSID), a large, nationally representative panel dataset and the Current Population Survey (CPS), a large cross-sectional dataset (with a limited panel aspect) used, among other things, to construct the official unemployment rate numbers for the US. LaLonde (1986) combines the men from the dropout, ex-addict and ex-convict target groups into a single male group. His female group consists of the women from the AFDC target group. He then creates comparison groups from the PSID and CPS datasets 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.

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 to 1979 who were between 20 and 55 years old and did not report being retired in 1979.[7] This comparison group represents a random (putting aside issues of non-random survey response in the PSID) sample of a much broader population than that implicitly defined by the NSW AFDC women

[7] It is not clear in either the published LaLonde (1986) paper or the unpublished LaLonde (1984) working paper whether the age restriction is imposed in 1975, 1979, or both. We impose it in 1975.

eligibility criteria. The other three 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.[8]

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, it should unambiguously increase the fraction of women in PSID-2 who would have proven eligible for NSW relative to PSID-1.[9] 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 requirement of receiving AFDC in 30 of the previous 36 months imposed on the NSW participants.

LaLonde's (1986) PSID-3 comparison group takes the subset of PSID-2 sample observations not currently employed at the time of their 1976 interview. In our view, given that random assignment for the NSW AFDC women starts in January 1976, this restriction represents conditioning on an outcome, so we do not consider this group in our empirical work. His PSID-4 comparison group takes the subset of PSID-1 (not PSID-2 or PSID-3) respondents with youngest children no less than five years old.[10] This comparison group imposes a different aspect of the NSW AFDC women eligibility than does PSID-2. We view the AFDC receipt criterion as the substantively more important of the two and so do not analyze this comparison group below. It would be interesting to analyze a comparison group whose definition pushed the available PSID information as hard as possible to mimic the NSW eligibility rules in future work.

[8] Moreover, a PSID comparison sample that met all of the NSW AFDC women eligibility criteria would contain a very modest number of observations. Presumably, issues such as these motivated the (expensive) collection of a dedicated sample of eligible non-participants in the National Job Training Partnership Act Study.

[9] 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 non-recipients reporting receipt.

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

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. This represents a replication in the strict sense defined by Clemens (2015). Within the set of analyses that fall within his definition of a replication ours represents a particularly ambitious one. We do not simply check LaLonde's code (which is in fact not available). We do simply attempt to recreate his tables using his data set and our own code, as DW (1999) and ST (2005a) do. Rather, as noted in the introduction, we start from scratch and attempt to recreate his analysis files for the AFDC women in the NSW demonstration and the corresponding PSID comparison groups and then use our best shot at his analysis files to undertake (most of) the analyses presented in LaLonde (1986).[11]

The appendix provides a (very) detailed account of our efforts to recreate LaLonde's (1986) analysis files for the NSW and PSID from the raw data. We highlight only the major issues and basic patterns here.[12] For the NSW data, we match the sample sizes exactly and come very close in terms of matching means. 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.[13] In

[11] The CPS data that LaLonde (1986) used included matching administrative data on earnings. We have not been able to locate these matched data and would appreciate any pointers readers may have regarding their availability.
[12] As noted by Lalonde (1986), a problem with the NSW data is that the public file does not include the calendar date for any of its interviews. However, every survey includes the monthly unemployment rates at the participant's site during the second, fifth, and eight months prior to each interview. We replicate Lalonde's work by computing the month and year of the baseline for each experimental participant by matching their unemployment series with the one reported in various issues of *Employment and Earnings*. With that information, we calculate real earnings for each quarter before and after assignment.
[13] Nominal earnings were converted to real earnings using the monthly CPI-W reported in the Survey of Current Business. All real earnings are in 1982 dollars.

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

For the PSID comparison group, 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 679 observations compared to 595 in Table 2 of LaLonde (1986), and our PSID-2 has 188 compared to his 173, a smaller difference both absolutely and proportionally. We offer three main explanations for this difference: First, LaLonde used a different release of the PSID than we do. At the time he did his empirical work, only the "Public Release 1" version of the data were available. At the time of our replication, we could gain access only to the more recent "Public Release 2" (or "final release") version. A FAQ on the PSID web page states that "The term "Public Release II" was previously used to refer to files which had undergone additional data checks to correct a very small number of cases";[14] this does not sound like enough in the way of modifications to generate our non-trivial sample size difference, but we have not managed to locate more detail about the specific nature of the changes from one version to the next to entirely rule this explanation out and the right sort of change in the coding of, say, the headship or retirement variables could certainly have the effect of enlarging our sample.

The second potential explanation consists of an error on LaLonde's part. As we do not have LaLonde's code, we can only address this explanation indirectly. We do have LaLonde's sample of men and the results of our attempt to replicate his PSID comparison groups for men. Our analysis of the data on men suggests that LaLonde's (1986) PSID-1 comparison group for men contains a number of spurious duplicate observations. Of course, our problem in regard to the data on women concerns not too many observations but rather too few, which leads us to

[14] https://psidonline.isr.umich.edu/Guide/FAQ.aspx?Type=ALL#30

think that whatever led to the error in LaLonde's (1986) PSID comparison groups for men did not lead to the same error in his comparison groups for women.

Rather obviously, the third potential explanation consists of an error on our part. We have checked and re-checked our code multiple times. We remain somewhat uncertain about exactly which PSID variable and responses LaLonde (1986) uses to define "retirement", but reasonable alternatives that we considered did not produce samples noticeably more similar to LaLonde's. We also plan to make available our code so that others can check what we did. The appendix provides yet more detail about our replication of LaLonde's (1986) PSID comparison samples for the women.

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 and using our samples. Our somewhat larger samples also have somewhat higher mean earnings. For example, in 1979, his PSID-1 has a mean of $8016 while ours has a mean of $8892; similarly for the PSID-2 his sample has a mean of $3569 and ours has a mean of $4641. Consistent with earnings having a relatively high variance in these populations, and the not-so-very-large samples, the differences vary from year to year and from sample to sample. Taking note of these differences, we proceed with the remainder of our analysis.

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 allows 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, employment in 1976, and number of children, along with receipt of AFDC in some specifications.[15] We do not consider the third specification because employment in 1976 represents an outcome for some observations randomly assigned early in 1976. LaLonde (1986) does not explicitly make a case that these conditioning variables suffice for exogeneity, other than noting that the literature as of his writing (as it does now) emphasizes the value of conditioning on pre-program outcomes, which suggests, quite reasonably, 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 non-linearities 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).

Finally, LaLonde (1986) applies the Heckman (1978) two-step estimator for the (at the time commonplace) bivariate normal selection model, using various (and no) exclusion

[15] We say "apparently" because the published paper never makes this covariate set explicit.

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, over-represents NSW participants relative to the population.[16] Second, as pointed out in ST (2005a), who learned it from Stata's `probit` command, which helpfully checks for such things, one of the exclusion restrictions employed to identify the model, residence in a Standard Metropolitan Statistical Area (SMSA), is a perfect one-way predictor of treatment status. The remaining exclusion restrictions – 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.

Table 3A presents the estimates from LaLonde's (1986) Table 4 based on the NSW women and the PSID-1 and PSID-2 comparison groups. Table 3B presents the corresponding estimates using our versions of the NSW AFDC women and the corresponding PSID-1 and PSID-2 comparison groups. Table 3B presents both experimental and non-experimental impact estimates, which the reader should compare to one another, and non-experimental bias estimates, which the reader should compare to zero.

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 very modestly between LaLonde's PSID comparison groups and our versions. For example, we see that the differences for PSID-1

[16] 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.

equal -$6,443 in LaLonde (1986) and -$6,707 in column (2) of Table 3B. 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 estimator in columns (6) and (7) and the selection-on-observed variables estimator that includes pre-program earnings in columns (8) and (9), we find substantially smaller biases than LaLonde (1986). In column (7), for the PSID-2 comparison group, LaLonde (1986) finds a difference of (2392 – 883) = $1509, compared to (1337-839) = $498 and $522 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). Here the biases turn out quite low, less than $200, for both the PSID-1 and PSID-2 compared to both the experimental treatment group and the experimental control group. 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 very low biases run against the claims in ST (2005) regarding the importance of time-invariant differences due to geography and/or earnings measurement between the NSW and the PSID.

5. Replicating Dehejia and Wahba (1999, 2002)

The Dehejia and Wahba (1999, 2002) [hereinafter DW] papers innovate in four main ways relative to LaLonde (1986): 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, DW assume the validity of a particular identification strategy and examine the performance of alternative econometric

estimators that build on that strategy. In particular, they assume that the variables at hand suffice to make an assumption of "selection on observed variables" plausible and investigate alternative estimators all of which build on that assumption.[17] Second, they investigate several estimators not previously applied to the data from LaLonde (1986). Though relatively common in the applied statistics literature at the time, the estimators they consider were not at all familiar in the empirical economics literature. Third, they emphasize the common support, or overlap, issue, and show its empirical relevance in the context of LaLonde's data on men. Finally, they trim LaLonde's sample of men in order to better justify the conditional independence assumption. We discuss these contributions in more detail in the remainder of this section, and explore them in the context of the NSW women. In the terminology of Clemens (2015), this section represents an extension of LaLonde (1986) to new estimators and an extension of DW (1999, 2002) to a different subset of the populations treated in the NSW demonstration.

DW (1999, 2002) did a very reasonable thing in applying matching estimators to the data from LaLonde's (1986) paper. His setup provides a context wherein we would expect matching estimators to make a difference relative to parametric linear models with only main effects included. As shown in DW (1999, 2002) and ST (2005a), in both the PSID-1 comparison group and the CPS-1 comparison group (not examined here for reasons noted above) a large fraction of the observations look nothing at all like anyone in the NSW, with the result that they have estimated propensity scores very close to zero. These observations play no role in the matching estimates but potentially play a very large role in determining the coefficients in a parametric linear model. A linear model that fits well in the regions of the data rich in these incomparable comparison group observations may not fit well in the region of the data containing the NSW

[17] In fact, the assumptions underlying matching and parametric linear regression differ slightly, with the matching assumptions slightly weaker; see e.g. Frölich (2008) for discussion.

observations, thereby yielding bias in the parametric estimates that the matching estimators avoid by assigning zero (or very low, depending on the particular estimator) weight to the incomparable comparison group units.

As described in detail in Smith and Todd (2005a) DW define their subsample of men based on two variables: date of random assignment and the value of earnings in "1974". In particular, they take observations with non-zero earnings in "1974" only if randomly assigned in January through April 1976. They do this in order to focus on a sub-population for which they have (more or less) two years of pre-random-assignment earnings. The literature, both early, as with Ashenfelter (1978) and Ashenfelter and Card (1985) or more recent, as with HIST (1998) or Andersson, Holzer, Lane, Rosenbaum and Smith (2013), clearly signals the importance of conditioning on a relatively rich set of pre-program outcomes, as these proxy, at least in part, for many otherwise unobserved variables that affect both program participation and outcomes in the absence of program participation. The asymmetric handling of those with zero earnings in months 13-24 before random assignment (what they call 1974 and we, following ST (2005), call "1974") presumes, again not unreasonably, a greater temporal stability in earnings among this group.

Because DW did not consider the women in their paper for the reasons explained above, we cannot know exactly what they would have done with this sample. Imposing the same rule that DW use on the men captures a grand total of only 12 women with non-zero earnings in "1974". We thus expand the sample to include NSW women with non-zero earnings in "1974" randomly assigned anytime in 1976. Table 4, inspired by Table 2 in Smith and Todd (2005a), graphically illustrates our sample definition. We label the resulting sample the "DW" sample

throughout our analysis; readers keen to assign praise (or blame) should keep in mind that, unlike the DW sample in Smith and Todd (2005a), DW inspired this sample but did not choose it.

Table 5 provides descriptive statistics for the DW sample of the NSW AFDC women. They reveal a DW 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 DW 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 DW sample. Somewhat surprisingly, this difference largely, but not entirely, disappears in the post-random-assignment period.

As laid out in ST (2005), all matching estimators[18] have the basic form

$$\Delta^M = \frac{1}{n_1} \sum_{i \in \{D_i = 1\}} \left[Y_{1i} - \sum_{j \in \{D_j = 0\}} w(i, j) Y_{0j} \right]. \quad (1)$$[19]

The potential outcomes notation remains as above and *D* again indicates wanting to participate in NSW, defined as undergoing random assignment. Thus $i \in \{D_i = 1\}$ indicates either the NSW treatment group or the NSW control group, depending on whether we seek to estimate the treatment effect or the bias non-experimentally; n_1 indicates the number of units in the corresponding set. Finally, $w(i, j)$ indicates the weight that comparison group observation "*j*" receives in the construction of the estimated expected counterfactual outcome for experimental observation "*i*". Only the weights $w(i, j)$ differ among the multitudinous variants of matching

[18] Following e.g. Heckman, Ichimura, Smith and Todd (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 can single nearest neighbor matching.

[19] This formula applies to the ATET. The corresponding formula for the ATE is straightforward.

now available in the literature; put differently, each different matching estimator defines an algorithm for constructing the weights $w(i, j)$.

Over time some of the applied econometric literature (though not the parallel literature in statistics) has come to think of propensity score matching as an application of non-parametric regression. In this interpretation, the second term in square brackets in equation (1) is the predicted value from a non-parametric regression of the untreated outcome Y_0 on the estimated propensity score $\hat{P}(X)$ where the weights then depend on the particular smoother used in the non-parametric regression. Thinking about the problem in this way allows the researcher to draw on the large technical literatures on theoretical and applied non-parametric regression, usefully summarized in e.g. Pagan and Ullah (1999) and Li and Racine (2006); thinking about the problem in traditional case-control terms masks this important conceptual connection.

Following DW, we apply three variants of matching in this section: propensity score stratification, single nearest neighbor matching with replacement, and weighted least squares using weights from single nearest neighbor matching with replacement. Propensity score stratification, described in Rosenbaum and Rubin (1984) and somewhat popular in the applied statistics literature, defines intervals of the estimated propensity score. Within each interval, the mean of the comparison group units' untreated outcomes serves as the second term in equation (1) for every treated unit in the interval. Two schools of thought characterize the implementation of this estimator. One school holds the number of strata fixed and augments the propensity score model to achieve balance in the estimated scores within strata. DW (1999) adopts this method. The second method holds the propensity score specification fixed and increases the number of strata until it achieves within-stratum balance in the estimated scores. We adopt the latter approach, starting with 10 strata with borders defined by deciles of the pooled propensity score

distribution as in Plesca and Smith (2007). As it turns out, we did not require additional strata for either comparison group.

Nearest neighbor matching represents the oldest (and most literal) form of matching. In this algorithm, each experimental unit gets matched to the nearest comparison group unit based on some distance metric. In our case, we use absolute distances in the estimated propensity score to determine who is near and who is far, but the broader literature, particularly outside economics, often considers other distance metrics, such as the Mahalanobis distance. In terms of equation (1), single nearest neighbor matching implies $w(i, j) \in \{0,1\} \forall i, j$. Matching can proceed with different numbers of nearest neighbors (the tuning parameter choice in this context) and with or without replacement. Matching with replacement allows a given comparison unit to match to more than one experimental unit; matching without replacement forbids such promiscuity. DW (2002) provides a clear and compelling description of why matching with replacement makes sense in the context of the NSW men. More generally, matching with replacement reduces bias at the cost of increased variance; it particularly makes sense in contexts with relatively few comparison observations that "look like" the experimental observations. We follow DW in matching with replacement and defaulting to a single nearest neighbor.[20] As noted in Busso, DiNardo and McCrary (2014), single nearest neighbor matching with replacement represents a conservative choice that minimizes bias at the expense of additional variance.

DW (1999, Table 3) also combine single nearest neighbor matching with replacement with ex post adjustment via a parametric linear model that includes the same conditioning variables as the propensity score. Equivalently, they perform weighted least squares using the

[20] The exception to this, following ST (2005b), concerns ties, where we take the mean outcome of all tied observations as the "match".

weights from the matching. While this estimator does not fit directly into equation (1), the ex post regression can reduce both finite sample bias due to imbalances that linger after the matching and improve efficiency by sucking up residual variance. See e.g. Ho, Imai, King and Stuart (2007), who conceive of matching as a "pre-processor" and Abadie and Imbens (2011) for the formal econometrics.

We use a parametric propensity score model, specifically a logit. We include in the propensity score models in this section only the covariates considered in the DW 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 grab the efficiency gain noted by Smith and Todd (2005a,b) associated with using both the experimental treatment and control groups in the propensity score estimation throughout our analysis.

Following the literature, 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.[21] These tests mimic the balance tests typically done in random assignment studies. A modest literature considers alternative balance tests; see in particular Smith and Todd (2005b), Imai, King and Stuart (2008), and Lee (2013). Given our smallish samples we keep things relatively simple and use as our metric of balance the

[21] Promising to make our specification more flexible on those happy occasions when additional NSW and PSID observations appear transforms our semi-parametric procedure into a non-parametric one. Another interesting road to go down considers explicitly semi-parametric propensity scores, as in Lehrer and Kordas (2013).

standardized differences proposed in Rosenbaum and Rubin (1983) as implemented in the `pstest` package for Stata by Leuven and Sianesi (2003).[22]

In addition to a base specification incorporating each of the DW variables as a main effect we explored, guided in part by which variables proved recalcitrant in the balance tests and in part by intuition left over from the papers about the NSW men, a variety of other specifications as well. These added flexibility via squared terms in age and education, interactions between age and education and between education and race/ethnicity, interactions between marital status and race/ethnicity and squared and interaction terms in the earnings variables and related zero earnings indicators, both among themselves and with race/ethnicity. The many interactions with race/ethnicity grew out of difficulties in balancing the Hispanic indicator; ultimately we decided to worry less about it than about the other variables due to the small number Hispanic observations in the PSID-1 comparison group, just 11, reflecting the small Hispanic populations in the sites contributing to the NSW AFDC women target group.

The balance tests led us to a model that includes, in addition to main effects in all of the variables that DW 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, as well as interactions between earnings in "1974" and 1975 and between the indicators for zero earnings in "1974" and 1975. Tables containing average derivatives from the estimated propensity score models appear in the appendix. They contain no substantive surprises.

[22] We also take a rain check on the automated propensity score specification selection algorithm outlined in Section 13.3 of Imbens and Rubin (2015).

Before turning to our estimates using the estimators employed by DW applied to the NSW women and the PSID comparison group, we address three additional issues of implementation: choice-based sampling, comparison group contamination, and common support. By construction, we have a choice-based sample. It includes the NSW experimental population (putting aside survey non-response) and a random sample of the broader populations from which we draw the PSID-1 and PSID-2 comparison groups. But the relative proportions of the two samples in our analysis data do not match their relative proportions in the broader population; instead, our data wildly over-represent the NSW experimental units. We use the logit model for our propensity scores in part because, as is well known, it is robust to choice-based sampling in the sense that only the estimated intercept differs from what would be obtained under simple random sampling. This property, combined with the fact that we consider only the ATET (or the bias in the ATET) rather than the ATE means that our matching and weighting estimators remain consistent in the presence of the choice-based sampling.[23]

Comparison group contamination arises when members of the comparison group receive the treatment under study but the 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 populations from which we draw PSID-1 and PSID-2, contamination does not raise any substantive concerns.

In regard to common support, the literature typically (but often implicitly) assumes that the common support assumption holds in the population but imposes some additional, more restrictive version of common support in the sample, promising to relax these finite-sample

[23] To estimate the ATE requires the true population proportions, which then weight the estimates of the ATET and the ATNT. When estimating the ATET, we seek merely to reweight the comparison observations to match the distribution of conditioning variables in the experimental population, of which our data provides a consistent estimate. This does not require knowledge of the population proportions. See Heckman and Todd (2009) for a somewhat different take on the issue.

strictures as the sample becomes larger. In a spirit of approximate replication for our DW-inspired analyses, we drop treated observations with estimated propensity scores above the maximum or below the minimum of the estimated scores of the comparison group units.[24]

Figures 1A to 1D display the post-trimming 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 or DW) 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 and comparison group units. Most experimental observations have relatively high estimated propensity scores while most comparison units have relative low ones. Second, conditioning the comparison group on AFDC participation 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 DW (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 e.g. Crump, Hotz, Imbens and Mitnik (2009) for more discussion of common support issues.

Table 7 presents our estimates based on the estimators employed in DW; in particular, in Table 3 of DW (1999). Table 7A contains impact estimates obtained using an NSW treatment group and a comparison group. The reader should compare these to the appropriate experimental

[24] DW drop comparison units outside the interval defined by minimum and maximum of the estimated propensity scores of the treated units. We adopt the approach we do for the (not so attractive) reason that it is what `psmatch2` provides.

impact estimate, given in the table notes. Table 7B contains bias estimates obtained using an NSW control group and a comparison group. The reader should compare these estimates to zero. Within each table, the first and third rows correspond to the LaLonde NSW sample while the second and fourth rows correspond to the DW NSW sample; similarly, the first two rows in each table correspond to the PSID-1 comparison group and the second two rows correspond to the PSID-2 comparison group. Column (1) in each table gives the unconditional mean difference, column (2) the conditional (on main effects in each variable other than pre-period earnings) mean difference, column (3) the estimate from propensity score stratification, column (4) the estimate from single nearest neighbor matching with replacement, and column (5) the estimate from single nearest neighbor matching combined with regression adjustment.

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 7B, the bias falls (in absolute value) from -$4,237 to $984 with single nearest neighbor matching with replacement. Much (much) smaller differences emerge for the PSID-2 sample: again for the LaLonde sample in Table 7B the bias falls (in absolute value) from -$808 to $625 for the same estimator. The biases turn out similar in magnitude to the experimental impact estimates (show in the notes to Table 7); this makes them substantively moderate but still too large for one to want to rely on non-experimental evaluation in this context. They also have, as in DW (1999, 2002) and ST (2005a,b) relatively large sample standard errors, reflecting the smallish samples and the relatively large residual variance of earnings in this population.

Compared to the parametric linear regression estimators in Table 3, the matching estimators typically yield somewhat larger bias estimates. For example, the rich selection-on-observed variables estimator in column (9) of Table 3 shows a bias of -$166 with the PSID-1 and

-$196 with the PSID-2, in contrast to biases for the single nearest neighbor estimate with propensity scores based on the same covariate set of $984 and $625 for the PSID-1 and PSID-2 comparison groups, respectively.[25] As expected, the parametric estimators also yield noticeably smaller standard errors. The relative comparison of the matching and parametric linear regression estimators with the same covariate sets clearly differs from that found for the men in DW (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 when employing the PSID-2 comparison group.

Turning to secondary findings, we see that regression adjustment / bias correction makes little difference to the bias associated with the nearest neighbor matching estimates in our context and often leads to increased standard errors. As with the parametric estimates in Tables 3A and 3B, 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 biases in general for the DW-inspired sample than for the LaLonde sample, which differs strongly from the pattern found for the men by ST (2005a). 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.

6. Replicating Smith and Todd (2005a,b)

[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.

Smith and Todd (2005a, b) stand on the sturdy shoulders of LaLonde (1986) and DW (1999, 2002) and add to the within-study comparison literature (and, more narrowly, to the NSW within-study comparison literature) in several ways: First, they consider a second sub-sample of the LaLonde data that, like the DW sample, allows for conditioning on two years of pre-random-assignment earnings but does not treat those with earnings in months 13-24 asymmetrically based on whether those earnings equal zero or not. Second, they apply the difference-in-differences matching methods developed in HIST (1998). Third, they apply the kernel and local linear matching estimators developed in Heckman, Ichimura and Todd [hereinafter HIT] (1997, 1998). These estimators have important advantages relative to the propensity score stratification and nearest neighbor matching estimators applied in DW (1999, 2002). Fourth, they consider "pre-program" tests of over-identifying 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 DW specifications that pass their test fail the ST balance test. Sixth, following HIST (1998), they examine both non-experimental impact estimates, constructed using the experimental treatment group and the comparison group and non-experimental bias estimates, constructed using the experimental control group and the comparison group. Doing so fully exploits the information available in the data. Finally, they examine the sensitivity of the DW 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 seven contributions, we focus in this section on just three: the early random assignment sample, the alternative matching estimator, and difference-in-differences matching. We incorporate two more of the seven, namely presenting both non-experimental bias and

impact estimates and the various lessons learned from the sensitivity analyses, throughout the entire paper. We do not consider the ST (2005b) balancing test for the reasons outlined in the preceding section. We not consider the pre-program tests because we view them as less informative than ST (2005a) do.[26] The nature of our replication exercise relative to ST (2005a,b) parallels that for DW (1999, 2002); in the Clemens (2015) terminology, our work is an extension of theirs to a new demographic group subject to the same treatment at the same time about whom data were collected in the same way.

The ST (2005a) early random assignment sample for the NSW men includes individuals randomly assigned in January through April (inclusive) of 1976, i.e. during the first four months of random assignment. Their sample includes just 108 treated units and 142 control units, 78 and 118 fewer than the DW male samples, respectively. The benefit that ST (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 DW's (1999, 2002) valid concern about having two years of random assignment earnings to condition on, and do so (in some sense) even more strongly than they do.

In the context of the NSW women, we face a nasty tradeoff. 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 four months of 1976, we have only 66 treatment group observations and 64 control observations, small numbers indeed even by the low standards of the NSW literature. Thus, relative to ST (2005a) we trade some match quality

[26] In particular, because the propensity score specifications adopted in ST (2005a, b) 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.

between earnings in "1974" and 1974 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 through 1979. Relative to the LaLonde and DW 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 RA sample looks more like the LaLonde sample than the DW sample, a not surprising finding given serially correlated earnings and the fact that the DW sample omits individuals with non-zero earnings in months 13-24 prior to random assignment. For unknown reasons, the early RA experimental impact well exceeds that of either of the other two experimental samples.

Following HIST (1998) and HIT (1997, 1998), ST (2005a, b) consider alternative matching estimators in which the $w(i, j)$ in (1) come from kernel or local linear regressions of the untreated outcomes of the comparison group units on the estimated propensity scores. The kernel and local linear approaches have the advantage relative to single nearest neighbor matching that they make use of all comparison units similar to a given treatment unit (in the sense of having an estimated propensity score close in absolute value) rather than just the most similar one. This trades a modest increase in bias for a sometimes substantial decrease in variance. Local linear matching has the additional advantage relative to both kernel matching and nearest neighbor matching of reducing boundary bias. This bias arises for propensity score values near zero and one when the conditional mean function has a non-zero slope and results from the asymmetry in the density of observations around the evaluation point generated by the

boundary.[27] Because of its a priori advantages we present only local linear matching estimates here.[28]

In addition to local linear matching we also present estimates based on normalized inverse propensity weighting (IPW) in this section. For the treatment on the treated parameter, the normalized IPW estimator takes the form:

$$\hat{\Delta}_{TT} = \frac{1}{n_1}\sum_{i=1}^{n} Y_i D_i - \frac{1}{n_0}\sum_{i=1}^{n}\left(\frac{1}{n_0}\sum_{i=1}^{n}\frac{\hat{P}(X)(1-D_i)}{1-\hat{P}(X)}\right)^{-1}\frac{\hat{P}(X_i)Y_i(1-D_i)}{1-\hat{P}(X_i)}, \tag{2}$$

where the notation follows that used above with the addition of n_0, the number of treated units. The term in parentheses normalizes the weights to sum to one in the sample (as they do in expected value in the population).

Though not used in ST (2005a, b), looking at IPW follows in the spirit of ST's desire to examine estimators with a priori superior econometric properties than single nearest neighbor matching and propensity score stratification. The IPW estimator (in its non-normalized form) dates back to Horvitz and Thompson (1952). It has recently come to occupy an important place in the applied econometric treatment effects literature in a way that neither kernel matching nor local linear matching has managed to do.[29] We suspect this relative success in the literature results from IPW not requiring a (typically annoying) bandwidth choice and from its similarity to the popular methodology of DiNardo, Fortin and Lemieux (1996). IPW also, under certain

[27] Local linear matching is not costless: estimating a slope coefficient consumes a degree of freedom in every local regression and so increases variance. See the HIT (1997, 1998), HIST (1998) and ST (2005a, 316-317) for more on the technical details.

[28] We use the `psmatch2` implementation of local linear matching with an Epanechnikov kernel and a rule-of-thumb bandwidth based on the formula that minimizes the integrated mean squared of the estimated regression function. As noted in Frölich (2005) and Galdo, Black and Smith (2008), this is not in general the bandwidth that minimizes the mean squared error of the estimated treatment effect, which is the object of interest in our context. Both papers offer preferable alternative bandwidth selection schemes; we have not implemented either one in our analysis as our prior is that the large effort involved would not yield a corresponding benefit via improvements in our estimates.

[29] As a prosaic but practically consequential example, when Stata introduced their built-in treatment effect command `teffects` they included IPW but not kernel matching or local linear matching.

circumstances, attains the semi-parametric efficiency bound, as noted in Hirano, Imbens and Ridder (2003). In the normalized form that we employ, it generally performs well in the Monte Carlo analyses in Huber, Lechner and Wunsch (2013) and Busso, DiNardo and McCrary (2014). IPW has trouble empirically in cases with weak common support and with estimated propensity scores close to zero and one, where the latter leads to (always problematic) division by numbers close to zero. To deal with these issues, the literature recommends doing some trimming; we trim the two percent of the treated observations corresponding to the comparison group observations with the lowest estimated propensity score densities.

Tables 8A and 8B present the estimates. The basic format repeats that of Tables 7A and 7B, with non-experimental impact estimates in Table 8A and non-experimental bias estimates in Table 8B and the experimental impacts in the table notes. Each panel includes an additional row for the "Early RA" sample in addition to the LaLonde and DW samples. The columns present unconditional mean differences (1), estimates from single nearest neighbor matching with replacement (2), estimates from single nearest neighbor matching with replacement combined with ex post regression (3), IPW (4) and local linear regression matching (5).

The values in columns (2) and (3) differ from the corresponding elements in Tables 7A and 7B due to the different trimming rule applied in Tables 8A and 8B; following ST (2005a), we trim the experimental observations corresponding to the comparison units with the lowest two percent of estimated propensity score densities while Tables 7A and 7B followed (roughly) the simpler-to-implement scheme in DW (1999, 2002). The trimming rule changes the estimates very little, particular the regression-adjusted estimates in column (3).

The two new econometric estimators we consider in this section have similarly limited effects on the estimated biases. In Table 7B for example, comparing IPW to regression-adjusted

nearest neighbor matching shows that sometimes the bias increases a bit, sometimes it decreases a bit and sometimes leaves it more or less the same. As expected though, IPW rather dramatically decreases the estimated standard errors by making more efficient use of the available data. For those concerned with MSE rather than just bias, this pattern reaffirms the general finding from the Monte Carlo literature that IPW dominates single NN. For those of dubious methodology who simply count the stars, IPW would yield a lot more of them than either NN estimator, had we included them in our table. The LLR estimator reduces the bias relative to IPW and the NN estimators for most combinations of NSW sample and comparison group, but modestly so, and at the cost of substantially increased standard errors. An MSE criterion would not push the researcher toward this estimator. We find this result a bit puzzling given the lack of a corresponding pattern in the standard errors in Table 5, Panel B of ST (2005a). There the LLR estimator typically has smaller standard errors than the single nearest neighbor matching (compare their column (4) to their columns (6) and (7)), despite the fact that ST (2005a) present bootstrap standard errors for the NN estimator which the Monte Carlo evidence in Abadie and Imbens (2006), the working paper version of Abadie and Imbens (2008), suggests are likely too small rather than too large.

Finally, we consistently find lower (in absolute value) biases in the estimates for the early random assignment sample than for either the DW or the LaLonde samples. This pattern diverges sharply from the parallel analysis in ST (2005a), which consistent found much larger biases for the early RA sample. There it seems that the early RA 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 observations. Large falls in earnings during the period prior to participation represent less of an issue for the NSW AFDC

population, and so we conjecture that the estimates reflect improvements from better temporal alignment of the pre-program earnings variables and the timing of random assignment.

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 exogeneous 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 non-random selection into treatment that depends on both observed covariates and time-invariant unobserved variables. Modified to include first differencing, equation (1) becomes

$$\Delta^{M} = \frac{1}{n_1} \sum_{i \in \{D_i=1\}} \left[(Y_{1it} - Y_{0it'}) - \sum_{j \in \{D_j=0\}} w(i, j)(Y_{0jt} - Y_{0jt'}) \right] \tag{2}$$

where *t* indicates an "after" period and t' indicates a "pre" period and where the conditioning again works through the weights and thereby through the propensity scores.

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 e.g. Andersson, Holzer, Lane, Rosenblum and Smith (2013), who obtain roughly the same estimates with both approaches in their non-experimental evaluation of the Workforce Investment Act. Imbens and Wooldridge (2009) and Chabé-Ferret (2014) provide further discussion; more research to determine the mapping from earnings and program participation processes in the choice of approach would have great value.

Tables 9A and 9B display the estimates based on difference-in-differences matching, with the estimators for each column defined as in Tables 8A and 8B. 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 Tables 8A and 8B with the cross-sectional estimators: not much of a systematic effect of IPW or LLR on the estimated bias but substantially smaller standard errors for IPW than the nearest neighbor estimators and substantially larger standard errors for the LLR matching. Similarly, we obtain the same pattern across samples, with smaller estimated biases for the early RA sample relative to the LaLonde and DW samples.

Juxtaposing the cross-sectional matching estimates in Table 8 and the difference-in-differences matching estimates in Table 9 reveals that the latter typically have about $100 less bias than the former. This suggests a relative 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. ST (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 ST (2005a) put on these explanations.

7. Conclusion

Our replication of LaLonde's (1986) analysis of the NSW AFDC women, and extension of the related analyses in DW (1999, 2002) and ST (2005a,b) 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 did not succeed in closely 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, required to pin down the exact source of the differences, but the updating of the PSID remains in our view the leading candidate. Our troubles highlight the potential value of keeping major (if obsolete) releases of important data sets available on ICPSR so as to enable replication, as well as the value of the more 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 distinctions among 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 datasets, 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 question.

In regard to the applied econometrics, we find that, like ST (2005a), alternative matching and weighting estimators do not deliver large differences in estimated biases. However, the IPW estimator not considered by ST (2005a), but very much in the spirit of their analysis of alternative estimators (to parametric linear regression, propensity score stratification, and single nearest neighbor matching) with a priori desirable econometric properties 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 draw four major conclusions. First, as Michael Lechner frequently reminds (one of) us, the NSW data has 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. We agree with him that the methodological literature in applied econometrics (and more recently in the causal part of applied statistics) should find an alternative, somewhat larger, canonical data set for examining the performance of estimators. And we remind the reader again that some of the patterns we identify lack statistical significance, though within that set we confine our remarks to those that show robustness across samples and (where relevant) estimators.

Second, it makes sense to impose program eligibility criteria in defining a comparison group. In our context, this means requiring comparison group members to have received AFDC in 1975. The PSID data lack the detail to allow a complete imposition of the eligibility rules (and would yield a quite small comparison sample if they did), but comparison of the results from the PSID-1 and PSID-2 comparison groups shows that the PSID-2 comparison group, which embodies the AFDC receipt requirement, poses an easier selection problem for our conditional independence and conditional bias stability assumptions to solve and thereby leads to reduced bias.

Third, the AFDC women pose an easier selection problem than 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 DW (1999, 2002) and ST (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 of 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 men but not for the women.

Fourth, and finally, we do not find large differences in bias estimates between the estimators based on the conditional independence assumption, e.g. the matching and IPW estimators, and the estimators based on the conditional bias stability assumption, e.g. difference-in-differences matching and IPW using the before-after difference as the dependent variable. ST (2005a) found large differences between the estimates built on these two identification strategies, with much lower bias for the difference-in-differences estimators, and 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 nationally representative 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.

References

Abadie, Alberto and Guido Imbens. 2006. “On the Failure of the Bootstrap for Matching Estimators.” NBER Technical Working Paper No. 325.

Abadie, Alberto and Guido Imbens. 2008. “On the Failure of the Bootstrap for Matching Estimators.” *Econometrica* 76(6): 1537-1557.

Abadie, Albert and Guido Imbens. 2011. “Bias-Corrected Matching Estimators for Average Treatment Effects.” *Journal of Business and Economic Statistics* 29(1): 1-11.

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.

Angrist, Joshua. 1998. “Estimating the Labor Market Impact on Voluntary Military Service Using Social Security Data on Military Applicants.” *Econometrica* 66: 249-288.

Ashenfelter, Orley. 1978. “Estimating the Effect of Training Programs on Earnings.” *Review of Economics and Statistics* 6: 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: 648-660.

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(5): 885-897.

Chabé-Ferret, Sylvain. 2014. “Symmetric Difference in Difference Dominates Matching in a Realistic Selection Model.” Unpublished manuscript, Toulouse School of Economics.

Clemens, Michael. 2015. “The Meaning of Failed Replications: A Review and Proposal.” IZA Discussion Paper No. 9000.

Couch, Kenneth. 1992. “New Evidence on the Long-Term Effects of Employment and Training Programs.” *Journal of Labor Economics* 10(4): 380-388.

Crump, Richard, V. Joseph Hotz, Guido Imbens and Oscar Mitnik. 2009. “Dealing with Limited Overlap in Estimation of Average Treatment Effects.” *Biometrika* 96(1): 187-199

Dehejia Rajeev and Sadek Wahba. 1999. “Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes. *Journal of the American Statistical Association* 94: 1053–1062.

Dehejia, Rajeev and Sadek Wahba. 2002. Propensity Score Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics* 84: 151–161..

DiNardo, John, Nicole Fortin, and Thomas Lemieux. 1996. “Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach.” *Econometrica* 64(5): 1001-1044.

Fraker, Thomas and Rebecca Maynard. 1987. “The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs.” *Journal of Human Resources* 22(2): 194-227.

Friedlander, Daniel and Philip Robins. 1995. “Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods.” *American Economic Review* 85(4): 923-937.

Frölich, Markus. 2005. “Matching Estimators and Optimal Bandwidth Choice.” *Statistics and Computing* 15: 197-215.

Frölich, Markus. 2008. “Parametric and Nonparametric Regression in the Presence of Endogenous Control Variables.” *International Statistical Review* 76(2): 214-227.

Galdo, Jose, Jeffrey Smith and Dan Black. 2008. “Bandwidth Selection and the Estimation of Treatment Effects with Non-Experimental Data.” *Annales d’Economie et Statistique* 91-92: 189-216.

Heckman, James. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica* 47: 153-161.

Heckman, James, Neil Hohmann, 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(2): 651-694.

Heckman, James and V. Joesph 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: 862-880.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. “Characterizing Selection Bias Using Experimental Data.” *Econometrica* 66(5): 1017-1098.

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(4): 605-654.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1998. “Matching as an Econometric Evaluation Estimator.” *Review of Economic Studies* 65(2): 261-294.

Heckman, James, Robert LaLonde, and Jeffrey Smith. 1999. “The Economics and Econometrics of Active Labor Market Programs.” In Orley Ashenfelter and David Card (eds.), *Handbook of Labor Economics, Volume 3A*. Amsterdam: North-Holland, 1865-2097.

Heckman, James, Lance Lochner and Petra Todd. 2007. “Earnings Functions, Rates of Return, and Treatment Effects: The Mincer Equation and Beyond.” In Eric Hanushek and Finis Welch eds. *Handbook of the Economics of Education Volume 1*. Amsterdam: North Holland. 307-458.
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(457): 313-348.

Heckman, James and Petra Todd. 2009. “A Note on Adapting Propensity Score Matching and Selection Models to Choice Based Samples.” *Econometrics Journal* 12(S1): S230-234.

Hirano, Keisuke, Guido Imbens, and Geert Ridder. 2003. “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score.” *Econometrica* 71(4): 1161-1189.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” *Political Analysis* 15: 199–236.

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

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(3): 455-477.

Horvitz, D. and D. Thompson. 1952. “A Generalization of Sampling Without Replacement from a Finite Universe.” *Journal of the American Statistical Association* 47(260): 663-685.

Huber, Martin, Michael Lechner, and Conny Wunsch. 2013. “The Performance of Estimators Based on the Propensity Score.” *Journal of Econometrics* 175: 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(Part 2): 481-502.

Imbens, Guido and Jeffrey Wooldridge. 2009. “Recent Developments in the Econometrics of Program Evaluation.” *Journal of Economic Literature* 47(1): 5-86.

Imbens, Guido and Donald Rubin. 2015. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Kemper, Peter, David Long, and Craig Thornton. 1981. *The Supported Work Evaluation: Final Cost Benefit Analysis*. New York: Manpower Demonstration Research Corporation.

LaLonde, Robert. 1986. “Evaluating the Econometric Evaluations of Training Programs Using Experimental Data.” *American Economic Review* 76: 604–620.

Lee, Wang-Sheng. 2013. “Propensity Score Matching and Variations on the Balancing Test.” *Empirical Economics* 44(1): 47-80.

Lehrer, Steven and Gregory Kordas. 2013. “Matching using Semiparametric Propensity Scores.” *Empirical Economics* 44(1): 13-45.

Leuven, Edwin and Barbara Sianesi. 2003. “PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing.” http://ideas.repec.org/c/boc/bocode/s432001.html. This version 4.2.1.

Li, Qi and Jeffrey Racine. 2006. *Nonparametric Econometrics: Theory and Practice*. Princeton, NJ: Princeton University Press.

Lise, Jeremy, Shannon Seitz and Jeffrey Smith. 2005. “Equilibrium Policy Experiments and the Evaluation of Social Programs.” NBER Working Paper No. 10283.

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.

Pagan, Adrian and Aman Ullah. 1999. *Nonparametric Econometrics*. New York: Cambridge University Press.

Plesca, Miana and Jeffrey Smith. 2007. “Evaluating Multi-Treatment Programs: Theory and Evidence from the U.S. Job Training Partnership Act” *Empirical Economics* 32(2-3): 491-528.

Rosenbaum Paul and Donald Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70: 41–55.

Rosenbaum, Paul and Donald Rubin. 1984. “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.” *Journal of the American Statistical Association* 79: 516–524

Smith, Jeffrey and Petra Todd. 2005a. “Does Matching Overcome LaLonde’s Critique of Nonexperimental Methods?” *Journal of Econometrics* 125(1-2): 305-353.

Smith, Jeffrey and Petra Todd. 2005b. “Rejoinder.” *Journal of Econometrics* 125(1-2): 365-375.

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(5): 1384-1417.

TABLE 1
Descriptive Statistics for AFDC Experimental and Comparison Group Samples

	Lalonde Sample		Calonico-Smith Sample			
					Comparison Group	
Variable	Treatments	Controls	Treatments	Controls	PSID-1	PSID-2
Age	33.37	33.63	33.33	33.46	37.02	34.38
	(7.43)	(7.18)	(7.52)	(7.57)	(10.62)	(9.38)
Years of School	10.3	10.27	10.27	10.27	11.30	10.47
	(1.92)	(2)	(2.03)	(2)	(2.78)	(2.11)
Proportion High School Dropouts	0.7	0.69	0.70	0.69	0.458	0.606
	(0.46)	(0.46)	(0.46)	(0.46)	(0.50)	(0.49)
Proportion Married	0.02	0.04	0.02	0.04	0.0191	0.0106
	(0.15)	(0.2)	(0.15)	(0.21)	(0.14)	(0.10)
Proportion Black	0.84	0.82	0.84	0.82	0.660	0.867
	(0.37)	(0.39)	(0.37)	(0.39)	(0.47)	(0.34)
Proportion Hispanic	0.12	0.13	0.12	0.13	0.0162	0.0213
	(0.32)	(0.33)	(0.32)	(0.33)	(0.13)	(0.15)
Month of Assignment (Jan. 78 = 0)	-12.26	-12.3	-12.23	-12.26		
	(4.3)	(4.23)	(4.39)	(4.4)		
Number of Observations	800	802	800	802	679	188

TABLE 2
Annual Earnings of NSW Treatments, Controls and PSID Comparison Groups

	Lalonde Sample				Calonico-Smith Sample			
			Comparison Group				Comparison Group	
Year	Treatments	Controls	PSID-1	PSID-2	Treatments	Controls	PSID-1	PSID-2
1975	895	877	7303	2327	862	879	7569	2239
	(81)	(90)	(317)	(286)	(82)	(91)	(295)	(261)
1976	1794	646	7442	2697	1783	618	7856	2955
	(99)	(63)	(327)	(317)	(95)	(59)	(305)	(312)
1977	6143	1518	7983	3219	6077	1502	8466	3573
	(140)	(112)	(335)	(376)	(139)	(111)	(313)	(378)
1978	4526	2885	8146	3636	4722	3212	8659	4050
	(270)	(244)	(339)	(421)	(247)	(267)	(319)	(408)
1979	4670	3819	8016	3569	4655	3833	8892	4641
	(226)	(208)	(334)	(381)	(227)	(208)	(335)	(503)
Number of Observations	600	585	595	173	600	585	679	188

Notes: Lalonde Sample constructed using only participants with valid earnings information in 1975 and 1979.

TABLE 3a
Earnings Comparisons and Estimated Training Effects for the NSW AFDC Participants using Comparison Groups from the PSID*

		NSW Treatment Earnings Less Comparison Group Earnings				Diff-in-Diff: Differences in Earnings Growth 1975-79 Treatments Less Comparisons		Unrestricted Diff-in-Diff: Quasi Difference in Earnings Growth 1975-79		Controlling for Observed Variables and Pre-Training Earnings	
		Pre-Training, 1975		Post-Training, 1979							
Comparison Group	Comparison Group Earnings Growth 1975-79	Unadjusted	Adjusted+	Unadjusted	Adjusted+	Without Age	With Age	Unadjusted	Adjusted+	Without AFDC	With AFDC
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Lalonde Sample											
Control	2942	-17	-22	851	861	833	883	843	864	854	-
	(220)	(122)	(122)	(307)	(306)	(323)	(323)	(308)	(306)	(312)	
PSID-1	713	-6443	-4882	-3357	-2143	3097	2657	1746	1354	1664	2097
	(210)	(326)	(336)	(403)	(425)	(317)	(333)	(357)	(380)	(409)	(491)
PSID-2	1242	-1467	-1515	1090	870	2568	2392	1764	1535	1826	-
	(314)	(216)	(224)	(468)	(484)	(473)	(481)	(472)	(487)	(537)	

* Each column presents estimated training effects for each econometric models and comparison group. The experimental mean impact estimate is $851. The first three columns present the difference between each comparison group's 1975 and 1979 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments. Standard errors in parentheses.
+ The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

TABLE 3b
Earnings Comparisons and Estimated Training Effects for the NSW AFDC Participants using Comparison Groups from the PSID*

Comparison Group	Comparison Group Earnings Growth 1975-79	NSW Treatment Earnings Less Comparison Group Earnings				Diff-in-Diff: Differences in Earnings Growth 1975-79 Treatments Less Comparisons		Unrestricted Diff-in-Diff: Quasi Difference in Earnings Growth 1975-79		Controlling for Observed Variables and Pre-Training Earnings	
		Pre-Training, 1975		Post-Training, 1979							
		Unadjusted	Adjusted+	Unadjusted	Adjusted+	Without Age	With Age	Unadjusted	Adjusted+	Without AFDC	With AFDC
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Calonico-Smith Sample											
Impact Estimates											
Control	2,954	-18	-22	821	841	839	839	824	824	861	-
	(220)	(122)	(122)	(308)	(307)	(324)	(324)	(308)	(308)	(307)	
PSID-1	1,323	-6,707	-4,927	-4,237	-2,856	2,470	2,244	944	731	695	646
	(241)	(324)	(332)	(415)	(437)	(340)	(343)	(384)	(386)	(183)	(179)
PSID-2	2,402	-1,418	-6,690	14	-146	1,391	1,388	645	635	497	-
	(461)	(213)	(329)	(493)	(508)	(497)	(497)	(498)	(498)	(231)	
Bias Estimates											
PSID-1	2,954	-6,690	-4,957	-5,059	-3,746	1,631	1,337	103	-166	614	573
	(220)	(329)	(337)	(410)	(428)	(330)	(332)	(370)	(371)	(179)	(176)
PSID-2	1,323	-1,360	-1,374	-808	-899	552	522	-162	-196	247	-
	(241)	(218)	(225)	(465)	(478)	(468)	(467)	(465)	(464)	(219)	

* Each column presents estimated training effects for each econometric models and comparison group. The experimental mean impact estimate is $821. The first three columns present the difference between each comparison group's 1975 and 1979 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments. Standard errors in parentheses.
+ The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

TABLE 4
Sample Composition

Month of Random Assignment	Zero Earnings in Months 13-24 Before RA	Non-Zero Earnings in Months 13-24 Before RA	Control	Treatment	Total
February-76	1	1	0	2	2
March-76	12	2	8	6	14
April-76	15	3	9	9	18
May-76	36	6	19	23	42
June-76	31	23	28	26	54
July-76	17	6	12	11	23
August-76	33	15	25	23	48
September-76	62	15	36	41	77
October-76	69	28	49	48	97
November-76	55	17	36	36	72
December-76	89	28	57	60	117
January-77	56	10	35	31	66
February-77	93	33	62	64	126
March-77	44	17	30	31	61
April-77	75	18	45	48	93
May-77	62	9	34	37	71
June-77	57	22	37	42	79
July-77	26	11	19	18	37
August-77	63	25	44	44	88
Total	896	289	585	600	1185

Early RA	A	564
DW	A+B	1040
Early Year 76	C	130
Early Year 77	D	496
Late RA	D+E	621

TABLE 5
Descriptive Statistics for Alternative AFDC Experimental Samples

	DW Sample		Early RA Sample	
Variable	Treatments	Controls	Treatments	Controls
Age	34.01	33.91	33.69	35.00
	(7.52)	(7.57)	(7.52)	(7.57)
Years of School	10.25	10.20	10.21	10.11
	(2.03)	(2)	(2.03)	(2)
Proportion High School Dropouts	0.71	0.69	0.72	0.71
	(0.46)	(0.46)	(0.46)	(0.46)
Proportion Married	0.02	0.04	0.02	0.04
	(0.15)	(0.21)	(0.15)	(0.21)
Proportion Black	0.83	0.80	0.87	0.86
	(0.37)	(0.39)	(0.37)	(0.39)
Proportion Hispanic	0.12	0.14	0.07	0.08
	(0.32)	(0.33)	(0.32)	(0.33)
Month of Assignment (Jan. 78 = 0)	-12.74	-12.68	-16.04	-16.04
	(4.39)	(4.4)	(4.39)	(4.4)
Number of Observations	526	514	285	279

TABLE 6
Annual Earnings of Alternative AFDC Experimental Samples

	DW Sample		Early RA Sample	
Year	Treatments	Controls	Treatments	Controls
1975	377	420	696	773
	(52)	(64)	(93)	(114)
1976	1703	399	3093	643
	(101)	(48)	(142)	(83)
1977	6126	1486	6677	1667
	(150)	(122)	(230)	(188)
1978	4514	3184	4182	2780
	(264)	(297)	(509)	(434)
1979	4589	3800	4847	3600
	(242)	(223)	(349)	(307)
Number of Observations	526	514	285	279

TABLE 7A

Impact Estimates Associated with Alternative Cross-Sectional Matching Estimators

Dependent Variable: Real Earnings in 1979

Sample	(1) Mean Diff.	(2) Adjusted Mean Diff.+	(3) Propensity Score Stratification*	(4) 1 N.N.*	(5) 1 N.N. Regression Adjusted*
	Comparison Group: PSID-1 Female Sample				
Lalonde	-4,237	669	1,257	1,804	1,783
Sample	(415)	(407)	(540)	(614)	(495)
DW	-4,303	1,132	1,734	1,545	1,599
Sample	(436)	(422)	(528)	(588)	(688)
	Comparison Group: PSID-2 Female Sample				
Lalonde	14	501	1,304	1,308	1,092
Sample	(493)	(517)	(770)	(786)	(820)
DW	-52	1,005	1,669	1,302	1,164
Sample	(505)	(543)	(592)	(607)	(955)

Notes: experimental mean impact estimates and associated standard errors (in brackets) are 821 (308) and 789 (330) for the Lalonde and DW samples, respectively.

+ Least squares regression: real earnings in 1979 on a constant, a treatment indicator, age, age2, education, no degree, black, Hispanic, real earnings 1974 and 1975.

* We discard the comparison units with an estimated propensity score less than the minimum (or greater than the maximum) estimated propensity score for treated units. Bootstrap standard errors in parentheses.

TABLE 7B

Bias Estimates Associated with Alternative Cross-Sectional Matching Estimators

Dependent Variable: Real Earnings in 1979

Sample	(1) Mean Diff.	(2) Adjusted Mean Diff.+	(3) Propensity Score Stratification*	(4) 1 N.N.*	(5) 1 N.N. Regression Adjusted*
	Comparison Group: PSID-1 Female Sample				
Lalonde	-5,059	-178	319	984	995
Sample	(410)	(391)	(538)	(661)	(472)
DW	-5,092	281	832	1,031	1,016
Sample	(431)	(411)	(528)	(589)	(714)
	Comparison Group: PSID-2 Female Sample				
Lalonde	-808	-224	470	625	621
Sample	(465)	(481)	(770)	(782)	(743)
DW	-841	166	950	724	729
Sample	(479)	(510)	(626)	(717)	(794)

Notes: under the conditional independence assumption the population value of the bias equals zero.

+ Least squares regression: real earnings in 1979 on a constant, a treatment indicator, age, age2, education, no degree, black, Hispanic, real earnings 1974 and 1975.

* We discard the comparison units with an estimated propensity score less than the minimum (or greater than the maximum) estimated propensity score for treated units. Bootstrap standard errors in parentheses.

TABLE 8A

Impact Estimates Associated with Alternative Cross-Sectional Matching and Weighting Estimators

Dependent Variable: Real Earnings in 1979

Sample	(1) Mean Diff.	(2) 1 N.N.*	(3) 1 N.N. Regression Adjusted*	(4) IPW*	(5) LLR Matching*
	Comparison Group: PSID-1 Female Sample				
Lalonde	-4,237	1,756	1,830	1,566	1,337
Sample	(415)	(642)	(608)	(336)	(2,067)
DW	-4,303	1,517	1,562	1,815	1,753
Sample	(436)	(619)	(622)	(340)	(723)
Early RA	-4,045	1,467	1,342	1,622	1,776
Sample	(565)	(721)	(1,000)	(398)	(872)
	Comparison Group: PSID-2 Female Sample				
Lalonde	14	1,251	1,393	1,466	1,069
Sample	(493)	(834)	(714)	(450)	(1,701)
DW	-52	1,277	1,326	1,798	1,442
Sample	(505)	(699)	(912)	(457)	(1,366)
Early RA	206	1,673	1,509	1,400	1,350
Sample	(593)	(921)	(863)	(608)	(1,384)

Notes: experimental mean impact estimates and associated standard errors (in brackets) are 821 (308), 789 (330) and 1247 (465) for the Lalonde, DW and Early RA samples, respectively.

* Trimming level for common support is two percent. Bootstrap standard errors in parentheses.

TABLE 8B

Bias Estimates Associated with Alternative Cross-Sectional Matching and Weighting Estimators

Dependent Variable: Real Earnings in 1979

Sample	(1) Mean Diff.	(2) 1 N.N.*	(3) 1 N.N. Regression Adjusted*	(4) IPW*	(5) LLR Matching*
	Comparison Group: PSID-1 Female Sample				
Lalonde	-5,059	971	1,164	744	414
Sample	(410)	(643)	(523)	(330)	(2,455)
DW	-5,092	1,020	1,213	1,026	890
Sample	(431)	(613)	(594)	(333)	(777)
Early RA	-5,292	761	928	376	393
Sample	(559)	(605)	(578)	(384)	(1,055)
	Comparison Group: PSID-2 Female Sample				
Lalonde	-808	580	786	644	515
Sample	(465)	(793)	(946)	(445)	(1,515)
DW	-841	871	853	1,009	705
Sample	(479)	(695)	(850)	(451)	(934)
Early RA	-1,041	130	-135	153	198
Sample	(557)	(892)	(689)	(591)	(4,567)

Notes: under the conditional independence assumption the population value of the bias equals zero.

* Trimming level for common support is two percent. Bootstrap standard errors in parentheses.

TABLE 9A

Impact 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	(1) Mean Diff.	(2) 1 N.N.*	(3) 1 N.N. Regression Adjusted*	(4) IPW*	(5) LLR Matching*
	Comparison Group: PSID-1 Female Sample				
Lalonde	2,470	1,565	1,604	1,403	1,182
Sample	(340)	(623)	(562)	(331)	(1,225)
DW	2,889	1,474	1,502	1,788	1,673
Sample	(348)	(599)	(830)	(335)	(646)
Early RA	2,828	1,243	1,113	1,570	1,532
Sample	(437)	(744)	(932)	(392)	(634)
	Comparison Group: PSID-2 Female Sample				
Lalonde	1,391	946	1,096	1,332	740
Sample	(497)	(847)	(741)	(443)	(2,221)
DW	1,810	1,195	1,232	1,749	1,334
Sample	(494)	(731)	(733)	(450)	(725)
Early RA	1,749	1,400	1,234	1,332	1,100
Sample	(577)	(879)	(772)	(598)	(878)

Notes: experimental mean impact estimates and associated standard errors (in brackets) are 821 (308), 789 (330) and 1247 (465) for the Lalonde, DW and Early RA samples, respectively.

* Trimming level for common support is two percent. Bootstrap standard errors in parentheses.

TABLE 9B

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	(1) Mean Diff.	(2) 1 N.N.*	(3) 1 N.N. Regression Adjusted*	(4) IPW*	(5) LLR Matching*
	Comparison Group: PSID-1 Female Sample				
Lalonde	1,631	815	957	564	325
Sample	(330)	(595)	(530)	(325)	(2,199)
DW	2,057	991	1,136	956	849
Sample	(341)	(608)	(678)	(330)	(946)
Early RA	1,504	658	742	245	260
Sample	(428)	(565)	(665)	(379)	(543)
	Comparison Group: PSID-2 Female Sample				
Lalonde	552	385	616	493	207
Sample	(468)	(754)	(709)	(437)	(656)
DW	978	773	739	917	621
Sample	(471)	(698)	(726)	(446)	(872)
Early RA	425	0	-200	7	15
Sample	(543)	(904)	(759)	(583)	(1,122)

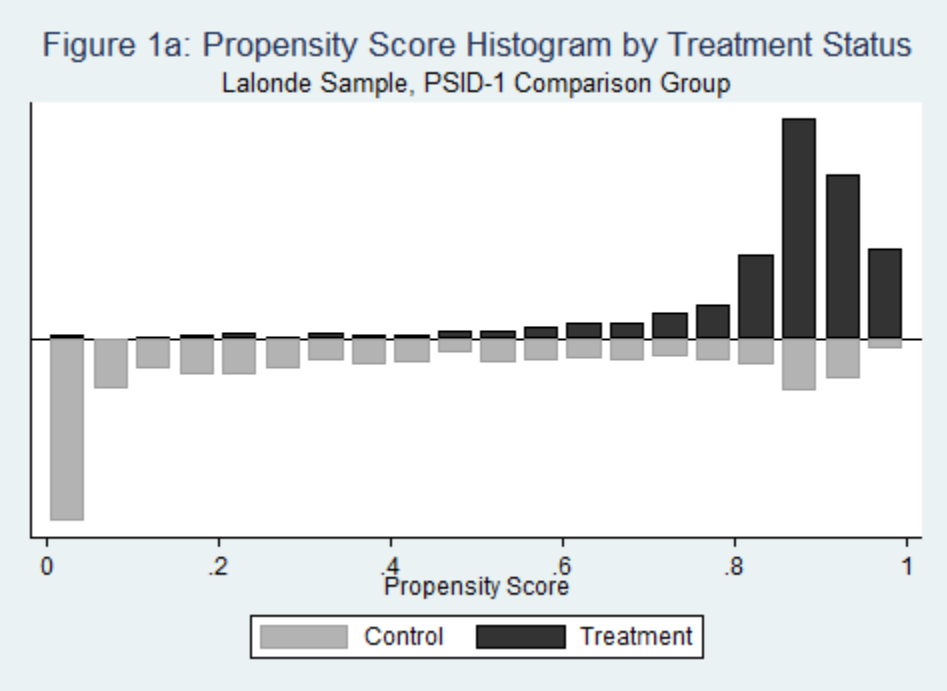
Notes: under the conditional independence assumption the population value of the bias equals zero.

* Trimming level for common support is two percent. Bootstrap standard errors in parentheses.

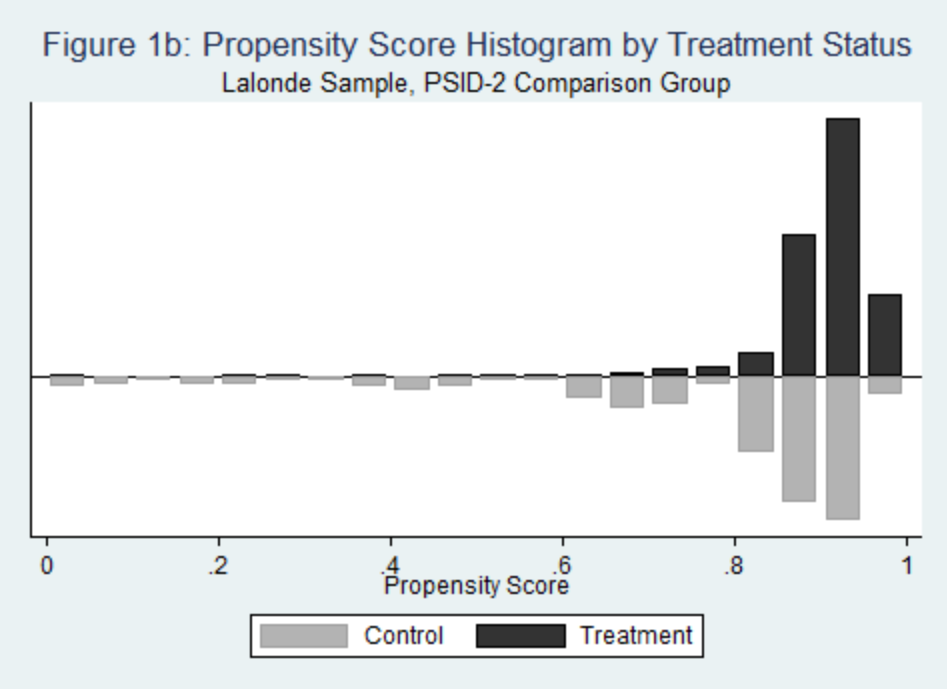
TABLE A1
Propensity Score Models
Average Derivate Estimates

VARIABLES	(1) Lalonde Sample		(2) DW Sample		(3) Early RA Sample	
	PSID-1	PSID-2	PSID-1	PSID-2	PSID-1	PSID-2
Age	-0.006***	-0.002	-0.005***	-0.002	-0.005***	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Schooling	0.000	0.003	-0.000	0.001	-0.002	-0.003
	(0.006)	(0.007)	(0.005)	(0.007)	(0.007)	(0.011)
High School Dropouts	0.029	0.031	0.011	0.021	0.021	0.036
	(0.022)	(0.023)	(0.022)	(0.025)	(0.030)	(0.040)
Married	0.065	0.069	0.073	0.074	0.115*	0.138
	(0.051)	(0.064)	(0.051)	(0.067)	(0.064)	(0.104)
Black	0.132***	0.055**	0.112***	0.050*	0.144***	0.075*
	(0.022)	(0.026)	(0.022)	(0.029)	(0.030)	(0.046)
Hispanic	0.314***	0.193***	0.265***	0.184***	0.266***	0.196**
	(0.050)	(0.056)	(0.049)	(0.059)	(0.066)	(0.093)
Real Earnings 1974	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Real Earnings 1975	-0.000**	-0.000	-0.000*	-0.000	-0.000*	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Zero Earnings 1974	0.905	0.877	1.077	0.962	1.236	1.400
	(24.360)	(35.066)	(38.674)	(26.557)	(54.070)	(52.440)
Zero Earnings 1975	-0.727	-0.422	-0.480	-0.324	-0.766	-0.640
	(19.968)	(18.585)	(20.059)	(11.046)	(33.038)	(24.087)
Observations	1,863	1,373	1,718	1,228	1,242	752

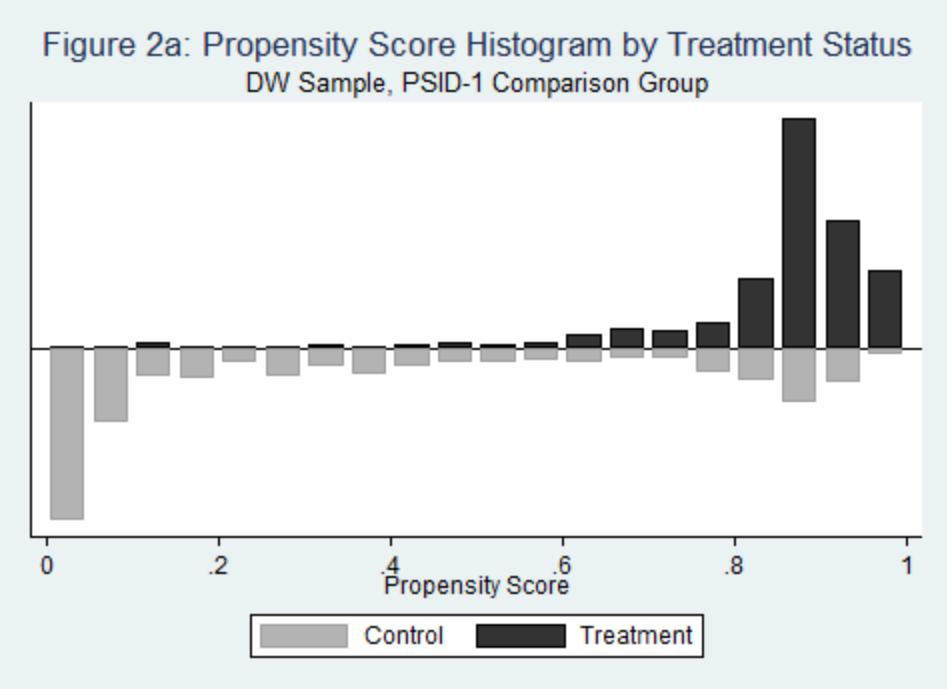
Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1



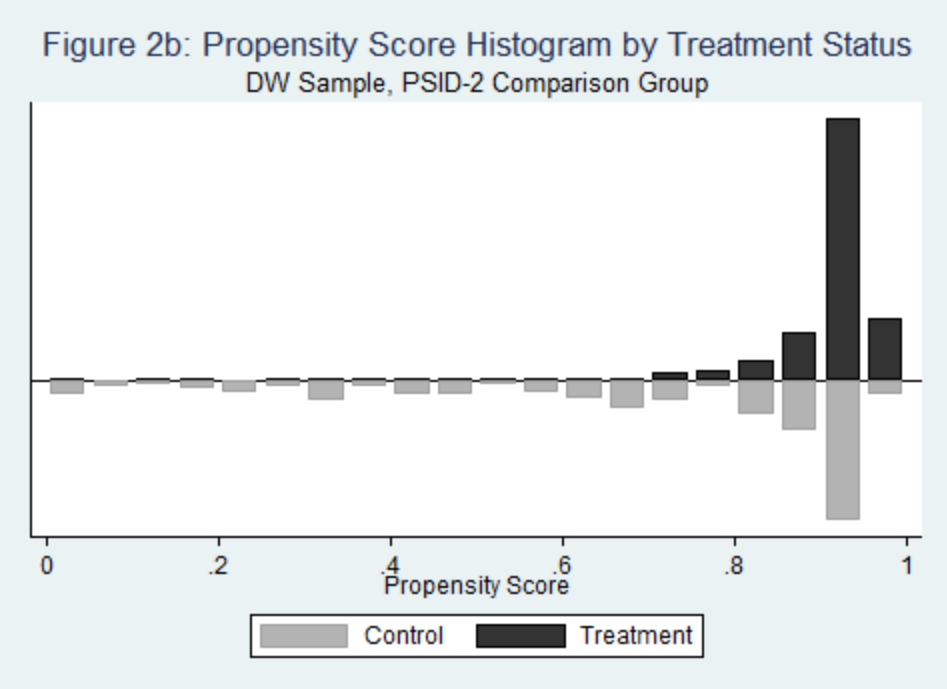
Notes: 148 PSID observations whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded.



Notes: 38 PSID observations whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded.



Notes: 197 PSID observations whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded.



Notes: 37 PSID observations whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded.

Appendix

Construction of the NSW samples

1. Download original file from: National Supported Work Evaluation Study, 1975-1979: Public Use Files, ICPSR 7865 http://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/7865.

2. Split the baseline, cross document and follow-up datasets using variable V0003:

0	Baseline
1	Cross Document Variables
9	1st follow up (9 months)
18	2nd follow up (18 months)
27	3rd follow up (27 months)
36	4th follow up (36 months)
37	4th follow up (37 months)

3. Using the monthly unemployment rates at the participant's site during the second, fifth, and eight month prior to each interview (variables V0956 V0957 V0958) we compute month and year of the baseline for each experimental participant by matching their unemployment series with the one reported in various issues of Employment and Earnings. The final output is the variable for month of assignment (moa), equal to zero for January 1978.

4. Compute real earnings for each quarter before and after assignment. Nominal earnings were converted to real earnings using the monthly CPI-W reported in the Survey of Current Business. All real earnings are in 1982 dollars.

Construction of the PSID samples

The PSID Sample is constructed using publicly available data files from the PSID web page http://psidonline.isr.umich.edu/. These data correspond to the PSID's "Public Release II", which

became available after LaLonde (1986) relied on the original "Public Release I" data. We have attempted to obtain the "Public Release I" data from the nice folks at the PSID but have been unsuccessful. Even if we were successful, it is not clear that we should focus our replication efforts upon it, as the PSID clearly views the "Public Release II" as superior. From another angle, using the "Public Release II" data changes the nature of our replication exercise, and does so in a way that we think makes it more interesting and more relevant.

In excruciating but useful detail, the construction of our PSID comparison sample proceeds according to the following steps:

1. Generate the main variables in the Family files by year (1975 to 1980) according to the scheme presented below.

2. Use the single file "1968-2005 Cross-Year Individual Files" to generate additional variables, especially the relationship of the respondent with the household head, one of the criteria used to select the final samples:

	1968	1975	1976	1977	1978	1979
ID	ER30001	ER30160	ER30188	ER30217	ER30246	ER30283
Relationship to HH Head		ER30162	ER30190	ER30219	ER30248	ER30285
Age		ER30163	ER30191	ER30220	ER30249	ER30286
Individual Weight		ER30187	ER30216	ER30245	ER30282	ER30312

3. We apply the following filters, as described in LaLonde (1986) to replicate the PSID-1 and PSID-2 comparison groups. We present these in terms of the underlying Stata code:

a. Keep only women

```
keep if ER32000==2
```

b. Keep only HH heads in every year:

```
keep if ER30162==1 & ER30190==1 & ER30219==1 & ER30248==1 &
ER30285==1
```

c. Age only individuals ages 20 to 55 in 1975

```
drop if age75>55
drop if age75<20
```

The last one is not explicitly mentioned in Lalonde's paper. Here we use the Moved In/Moved Out indicator to drop those heads that moved-in or out of the household any year from 1975:

```
drop if ER30193>0
drop if ER30222>0
drop if ER30251>0
```

To create PSID-2, we impose the additional condition that the respondent received AFDC in 1975. In terms of the underlying variables, this condition corresponds to:

```
keep if v5036 > 0
```

4. Merge the cross-year file with yearly Family files by year (1975-1980) using the ID variable in every year as the merging indicator (as recommended by the PSID staff).

Differences in descriptive statistics between the PSID comparison sample used by LaLonde (1986) and the one we created using the more recent PSID release are discussed in Section 3 of the main text.

Source Variables for PSID Replication Sample

Variable	1975	1976	1977	1978	1979	1980
ID 1968	V3909	V4423	V5336	V5835	V6446	V7050
ID 1975	V3802	V4430	V5343	V5842	V6453	V7057
ID 1976		V4302	V5344	V5843	V6454	V7058
ID 1977			V5202	V5844	V6455	V7059
ID 1978				V5702	V6456	V7060
ID 1979					V6302	V7061
ID 1980						V6902
Age	V3921	V4436	V5350	V5850	V6462	V7067
Gender	V3922	V4437	V5351	V5851	V6463	V7068
Years of Schooling	V4093	V4684	V5608	V6157		
High School Dropout	V4093<12	V4684<12				
Residence	V3806					
Number of Children	V3924					
Married	V4053==1	V4603==1	V5650==1	V6197==1		
Black	V4204==2	V5096==2	V5662==2	V6209==2		
Hispanic	V4204==3	V5096==3	V5662==3	V6209==3		
Age youngest child	V3925					
Employment Status	V3967	V4458	V5373			
Labor Income (previous calendar year)	V3858	V4373	V5283	V5782	V6391	V6981
AFDC recipient 1975		V5036>0				
Family Change Status		V4310	V5210	V5710	V6310	V6910
HH head move-in		V4312	V5212	V5712	V6312	V6912
HH head move-out		V4314	V5214	V5714	V6314	V6914
Family Weight		V5099	V5665	V6212		