Does Federally-Funded Job Training Work? Nonexperimental Estimates of WIA Training Impacts Using Longitudinal Data on Workers and Firms

Fredrik Andersson, Harry Holzer, Julia Lane, David Rosenblum, and Jeffrey Smith

June 7, 2022

Online Appendices

Online Appendix 1: Data and samples

LEHD

The Census Bureau established the Longitudinal Employer-Household Dynamics (LEHD) program in 1998 to provide information on employer and employee dynamics and the link between the two. The program draws on already existing survey and administrative data from both the demographic and economics directorates at the Bureau, and integrates them with Unemployment Insurance wage record data from its partner states.[1]

State Unemployment Insurance (UI) wage records sit at the core of the LEHD data. These records consist of quarterly reports filed by employers for each individual in covered employment, which includes roughly 96% of private non-farm wage and salary employment (Stevens 2007). The UI data provide less comprehensive coverage of agricultural employment and completely exclude federal government employees, self-employed individuals and independent contractors, and workers in the informal sector. According to US Department of Labor (1997), UI wage records measure "gross wages and salaries, bonuses, stock options, tips, and other gratuities, and the value of meals and lodging, where supplied." They do not include employer contributions to Social Security, health insurance, workers compensation,

[1] For more on the LEHD, see the Census program website: http://lehd.ces.census.gov/.

unemployment insurance, and private pension and welfare funds. Although UI administrative records contain some errors, staff at the LEHD program has invested substantial resources in cleaning the records and making them internally consistent; see e.g. Abowd et al. (2009).

The LEHD necessarily inherits some limitations from the underlying UI earnings data. Most states (including our two) do not require employers to report hours or weeks worked; as a result, we cannot measure hourly or weekly earnings and thus cannot easily distinguish between low wages and few hours worked as sources of low quarterly earnings. The data contain no information on employment separate from earnings; as such, we code employment in a calendar quarter as an indicator for non-zero earnings in a quarter, a process that will miss some extended leaves. In addition, for workers holding multiple jobs in a given calendar quarter the data provide no way to determine whether they hold them sequentially or in parallel. Finally, we have no direct information on why workers leave jobs, or on a range of personal characteristics (beyond basic demographics) typically captured in survey data. See Kornfeld and Bloom (1999), Hotz and Scholz (2002), Wallace and Haveman (2007) and Barnow and Greenberg (2015) for discussions comparing administrative and survey outcome measures.

The individual level data in the LEHD's Employment History File (EHF) is matched to the Census Numident data to obtain basic demographic information including date of birth, place of birth, sex and a crude measure of race and ethnicity. In similar fashion, we merge the WIASRD data with the EHF at the individual level. The LEHD is also matched to a number of other Census survey data sets, but because they have relatively small samples, and we have only two states, the intersection provides too little information for us to effectively utilize.

Using the registration date from the WIASRD data, we obtain quarterly earnings for the period from 12 quarters before WIA registration to 12 quarters after WIA registration for each registrant. We convert earnings to real 2020 dollars with the CPI-U-RS. In each quarter relative to WIA registration, we trim observations with earnings in the top half percent of the distribution separately for the treated and untreated observations.

The Employer Characteristics (ECF) file provides quarterly employment for each firm, from which we obtain our measure of firm size. The ECF also contains NAICS codes categorizing the industry of each firm at (roughly) the two-digit level. More precisely, we combine a few that are similar and frequently aggregated in applied work (e.g. 31-33 for manufacturing, 48-49 for transportation, 44-44 for retail trade). We use the industry information directly in the conditioning set for the last pre-WIA employer and we use it to construct our measure of industry change. Workers with no employer in the relevant period (i.e. the 12 quarters before WIA registration or the 12 quarters after WIA registration) have “unemployed” as their industry. To calculate turnover rates, we use the EHF dataset on an annual basis to find the number of accessions, separations, and end-of-year employment for each firm for each year. We define turnover as accessions plus separations divided by the sum of end-of-year employment in the previous year and the current year. Finally, our firm fixed effects come from Holzer et al. (2011).

After creating these firm characteristics, we set a worker’s pre-WIA firm characteristics equal to the characteristics of the worker’s last employer prior to WIA registration. For workers with more than one employer in their most recent quarter of employment prior to WIA registration we assign the characteristics associated with the employer with the highest earnings in the quarter. A similar process guides the creation of the post-WIA firm characteristics, which correspond to the last employer found in the 12 quarters after WIA registration.

The primary limitation of the employer data centers on the reporting unit. Although we often refer to the employer as a "firm," the actual reporting unit is an administrative, rather than an economic entity; in other words, the filing unit reflects an "Employer Identification Number," or EIN, rather than a specific establishment. This represents a distinction without a difference for the approximately 70% of workers with a single establishment employer, but for those who work for a multiple establishment employer, we cannot identify the exact employing establishment within the firm, and hence neither their industry nor their location. As noted above, we use the industry code imputed on the ECF for those cases; see Isenberg, Landivar, and Mezey for more about industry measurement in general and industry imputation in the ECF in particular.

Sample

Table A-1 describes the sample loss associated with the exclusions that led us from the raw data to our analysis sample. For State A, the sample loss associated with not having an exit date is the result of missing data on exit dates for registrants who leave before the end of our data rather than the result of individuals exiting after the end of our data.

Variables

Table A-2 provides detailed variable definitions for both the outcomes we consider and the conditioning variables we use. Table A-3 offers descriptive statistics. For completeness, it includes the variables discussed in the text whose descriptive statistics we also presented in Table 1.

Online Appendix 2: Estimators

As noted in the main text, we consider two econometric estimators that rely on our "selection on observed variables" assumption. We view these estimators as complements because of their contrasting costs and benefits. Inverse propensity weighting (IPW), despite an academic pedigree dating back to Horvitz and Thompson (1952), has become popular in the treatment effects literature only recently. Key papers in the literature are Hirano, Imbens and Ridder (2003) and Hirano and Imbens (2001). IPW has three primary benefits for our purposes: First, it does very well relative to other estimators in recent Monte Carlo horseraces run by Huber, Lechner and Wunsch (2013) and Busso, DiNardo and McCrary (2014). Second, it has the desirable asymptotic property that it attains the "semi-parametric efficiency bound", which means that (under certain conditions) it attains minimum asymptotic variance within the class of semi-parametric estimators; see Hirano, Imbens and Ridder (2003) on this point. Third, unlike many other semi-parametric treatment effects estimators, it does not require the choice of a bandwidth or other tuning parameter. In terms of our notation, the IPW estimator is given by

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

where n_0 denotes the number of untreated units and n_1 the number of treated units.

The downside to IPW lies in its sensitivity to estimated propensity scores close to one in finite samples. A quick look at the estimator makes the source of the trouble clear enough: the estimator divides by $1-\hat{P}(X)$. As a result, when the estimated propensity score lies very near one, small changes in its value can move the estimate a lot. As noted in Online Appendix 3, we have sufficient common support to avoid this problem. The Monte Carlo literature, in particular Busso, DiNardo and McCrary (2014), emphasizes the importance of normalizing the weights to sum to one in the sample for the finite sample performance of the estimator. The formula given here embodies that normalization. We obtain estimated standard errors using the non-parametric bootstrap.[2] Our bootstrap incorporates the estimation of the propensity scores.

We utilize single nearest neighbor matching on the estimated propensity score as our secondary estimator of choice. We do so despite its uniformly poor performance in terms of mean squared error in the Monte Carlo studies just cited, as well as in Frölich (2004). That poor performance in mean squared error terms masks a combination of quite good performance on bias, and truly awful performance on variance. The latter is perhaps not surprising given that nearest neighbor matching, particularly the most common variant with a single nearest neighbor used with replacement, completely ignores the information available in the data from comparison observations close to, but not closest to, particular treated units. In our view, the low bias, combined with its relative insensitivity to propensity scores close to one, makes nearest neighbor matching a good complement to IPW.

We present nearest neighbor matching estimates using one, three and five nearest neighbors as a sensitivity analysis. Increasing the number of neighbors trades off bias (due to using more distant matches) and variance (which decreases in the number of untreated units used to construct the

[2] We follow common practice in calling our populations "samples" and presenting standard errors and statistical tests consistent with this mislabeling; philosophically inclined readers should imagine meta-populations.

counterfactual). We chose these tuning parameters on a priori grounds rather than, say, via cross-validation as in Black and Smith (2004). Abadie and Imbens (2008) show that the bootstrap fails to provide consistent standard error estimates for the nearest neighbor estimator. Instead, we employ the consistent estimator in Abadie and Imbens (2016).

We use the same set of estimators when relying on the conditional bias stability assumption for identification, but instead of an outcome level as the dependent variable, we have a pre-post difference in outcomes as the dependent variable, as in Smith and Todd (2005).

Online Appendix 3: Determinants of training / estimated propensity scores

Table A-4 presents mean derivatives (a.k.a. average marginal effects) from our six models of the conditional probability of receiving training among WIA enrollees. Table A-5 presents the correlations of the estimated propensity scores from the six models. The pattern of correlations corresponds to the pattern of impacts; in particular, the scores from Models 3, 4, 5, and 6 all have correlations of at least 0.99. The similarity of the impact estimates they produce thus comes as no surprise.

Figures A-1a to A1-d display the distributions of estimated propensity scores from Model 6 separately by treatment status for the four combinations of state and funding stream. As expected, the treated units have higher estimated scores on average than the untreated units. In all four cases, the density for the untreated units reaches its maximum in the lower part of the support and the density for the treated units has at least a local maximum in the upper part of the support. The common support condition never comes close to failing. Though the estimated density is sometimes low for the untreated units in the upper part of the support, our relatively large sample sizes mean that we have enough comparison units for the estimator to perform well, as indicated by the estimated standard errors.[3]

Tables A-6a to A-6d offer standard balance statistics for all six models for all four combinations of state and funding stream. In particular, for each model we present two columns of standardized differences and two columns of variance ratios. In each case, the first of the two columns refers to the unweighted data and the second of the two columns refers to the weighted data, where in each case we weight the untreated data by the estimated propensity scores using the inverse propensity weights. Ideally, weighting moves the standardized difference to zero and the variance ratio to one. Rosenbaum and Rubin (1985) suggest a threshold of 0.2 for the reweighted standardized difference. Our models never cross this threshold and, indeed, only cross a lower threshold of 0.1 a handful of times. This relatively strong balance performance, combined with our time-limited access to the data, led us to proceed with propensity score models that include only main effects in the conditioning variables.

Online Appendix 4: Nearest neighbor matching estimates

Table A-7 compares our preferred IPW estimates to estimates obtained using nearest neighbor matching with replacement on the estimated propensity score with one, three and five nearest neighbors. As expected, the nearest neighbor estimates have larger standard errors than the IPW estimates, with the standard errors declining (up to noise) as the number of neighbors increases. The impact estimates tell the same qualitative story as our preferred IPW estimates in all cases.

[3] We promised to keep the exact bandwidth used for these figures a secret.

We do not find this particularly surprising given our relatively large sample sizes and the absence of serious support issues.

Online Appendix 5: Omitting "pre" period labor market outcomes from the conditioning

Tables A-8a and A-8b parallel Table 8 in the main text, but with propensity scores that omit all pre-program labor market outcomes, in order to examine the extent to which the differencing, plus conditioning on geography (in most models) and demographics, accomplishes the same thing. For State A, the estimates land pretty close to those in Table 8 for the adults, but are quantitatively different for the dislocated workers. In State B, the estimates are quantitatively different in all cases. In general, the difference-in-differences estimates that use the scores that do not include the pre-program outcomes yield more negative or less positive impact estimates, which means that they do not change the cost-benefit story for the dislocated workers.

Online Appendix References

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

Abadie, Alberto and Guido Imbens. 2016. "Matching on the Estimated Propensity Score." *Econometrica* 84(2): 781-807.

Abowd, John, Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin McKinney, Marc Roemer, and Simon Woodcock. 2009. "The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators." In Timothy Dunne, J. Bradford Jensen and Mark Roberts eds. *Producer Dynamics: New Evidence from Micro Data*. Chicago: University of Chicago Press. 149-230.

Barnow, Burt and David Greenberg. 2015. "Do Estimated Impacts on Earnings Depend on the Source of the Data Used to Measure Them? Evidence From Previous Social Experiments." *Evaluation Review* 39(2): 179-228.

Black, Daniel and Jeffrey Smith. 2004. "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching," *Journal of Econometrics* 121(1): 99-124.

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.

Frölich, Markus. 2004. "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators." *Review of Economics and Statistics* 86(1): 77-90.

Hirano, Keisuke and Guido Imbens. 2001. "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services and Outcomes Research Methodology* 2: 259-278.

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

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

Hotz, V. Joseph and Karl Scholz. 2002. "Measuring Employment and Income Outcomes for Low-Income Populations with Administrative and Survey Data." In *Studies of Welfare Populations: Data Collection and Research Issues*. National Research Council: National Academy Press. 275-315.

Huber, Martin, Michael Lechner and Conny Wunsch. 2013. "The Performance of Estimators Based on the Propensity Score." *Journal of Econometrics* 175:1-21.

Isenberg, Emily, Liana Landivar, and Esther Mezey. 2013. "A Comparison of Person-Reported Industry to Employer-Reported Industry in Survey and Administrative Data." U.S. Census Bureau, Social, Economic, and Housing Statistics Division Working Paper Number 2013-24.

Kornfeld, Robert and Howard Bloom. 1999. "Measuring Program Impacts on Earnings and Employment: Do Unemployment Insurance Wage Reports from Employers Agree with Surveys of Individuals? *Journal of Labor Economics* 17: 168–197.

Rosenbaum, Paul and Donald Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *American Statistician* 39(1): 33-38.

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

Stevens, David. 2007. "Employment That Is Not Covered by State Unemployment Insurance Laws." U.S. Census Bureau (Washington, DC) LEHD Technical Paper No. TP-2007-04.

US Department of Labor. Bureau of Labor Statistics. 1997. *Bureau of Labor Statistics Handbook of Methods*. Washington DC: Government Printing Office.

Wallace, Geoffrey and Robert Haveman. 2007. "The Implications of Differences between Employer and Worker Employment/Earnings Reports for Policy Evaluation." *Journal of Policy Analysis and Management* 26 (4): 737-753.