Preliminary WIA Net Impact Estimates: Administrative Records Opportunities and Limitations

by

Kevin Hollenbeck
Christopher T. King
Daniel Schroeder

New Tools for a New Era! Symposium
Bureau of Labor Statistics & the Workforce Information Council
Embassy Row Hilton Hotel
Washington, D.C.

July 23-24, 2003

This paper was prepared as part of the Administrative Data Research and Evaluation Project under agreement K-6558-8-00-80-60 for the Division of Research and Demonstration, Office of Policy and Research, Employment and Training Administration, U.S. Department of Labor. The authors acknowledge helpful comments and suggestions provided by Dan Ryan and Jeff Smith on earlier versions of the paper. Additional comments should be provided to Kevin Hollenbeck, W.E. Upjohn Institute for Employment Research (hollenbeck@upjohninstitute.org, 269-385-0431), and Christopher T. King and Daniel Schroeder, Ray Marshall Center for the Study of Human Resources, The University of Texas at Austin (ctking@uts.cc.utexas.edu, 512-471-2186; schroed@uts.cc.utexas.edu, 512-471-2196).

This paper presents preliminary results from an exploratory study estimating quasi-experimental net impacts of training services provided under the Workforce Investment Act (WIA) of 1998 on the employment and earnings of participants in seven states.[1] The study specifically focuses on individuals who exited WIA in Program Year (PY) 2000 (July 1, 2000 – June 30, 2001), comparing their labor market experiences during the first four quarters after exit to those of comparable individuals who were registered for WIA but did not receive training services.[2] This study exploits the availability of administrative data on participants who received core, intensive, or training services provided under WIA and, in one state, on registrants for Employment Services (ES) provided under the Wagner-Peyser Act. We have linked program and Unemployment Insurance (UI) wage record data from the seven states that are currently participating in the Administrative Data Research and Evaluation (ADARE) Project: Florida, Georgia, Illinois, Maryland, Missouri, Texas, and Washington. The ADARE Project, a collaboration of university and nonprofit organization partners, is funded by the U.S. Department of Labor's Employment and Training Administration.

We begin the paper with some background on quasi-experimental impact estimation for employment and training programs. We explain the approach we have adopted for estimating WIA net impacts on employment and earnings, two of the main outcomes of interest. We then provide our preliminary impact results, keeping the identity of the individual states confidential for now. We offer a brief discussion of our results and their implications before outlining next steps for our research. We conclude by examining the opportunities and limitations associated with the use of administrative records for net impact estimation for workforce development programs. This paper is still a work-in-progress.

[1] Prior versions of this paper focused on receipt of WIA training *or* intensive services as the treatment, but that approach was abandoned due to small numbers of potential comparison group members.
[2] In one of the states, we have compared outcomes for individuals who received *any* services through WIA (core, intensive, or training) to comparable individuals who registered with the Employment Service, but did not receive WIA services.

Background

The ideal approach for evaluating the effects of workforce interventions such as training services under WIA is an experimental one with random assignment of participants into treatment and control groups, typically with further statistical adjustments to account for any remaining treatment/control differences. Absent an experiment, a quasi-experimental approach may provide the best available estimates of program net impact.[3] This method uses comparison groups of individuals *similar* to those in the treatment group rather than random assignment. Quasi-experimental techniques have been used to evaluate state workforce training programs in Missouri (Mueser et al. 2003), Utah (Bowman, 1993) and Washington State (Hollenbeck et al., 2001, 2002), as well as welfare employment programs in Texas (King et al., 1994). Access to extensive longitudinal program administrative data greatly facilitates the ability to estimate net impacts through a quasi-experimental approach.

Until the mid-1980s, training program evaluations typically relied on quasi-experimental approaches, producing impact estimates that varied widely and were quite sensitive to the particular outcome measures used.[4] There were growing concerns about the reliability and validity of impact results for programs under the Comprehensive Employment and Training Act (CETA) based on quasi-experimental methods and their potential application to its replacement, the Job Training Partnership Act (JTPA). Among the problems these methods faced was the lack of local identifiers in the available data sources that would enable comparison groups to be constructed from the same local labor markets. In 1985, the Job Training Longitudinal Survey Advisory Panel (1985) chaired by Ernst Stromsdorfer advised the U.S. Department of Labor to jettison its plans for further quasi-experimental estimation and, instead, to devote its scarce resources to conducting a large experimental evaluation of JTPA. USDOL subsequently followed this advice, launching the National JTPA Study in 1986. Applicants were randomly assigned to treatment and control groups in three broad service strategies in 16 sites around the

[3] While some researchers prefer the term "non-experimental" to describe such approaches, we feel that "quasi-experimental" is more appropriate, in that it better captures the intent of these methodologies, that is, approximating the results of an experiment.
[4] Barnow (1987) and Friedlander et al. (1997, 2000) review this literature.

country as part of this successful experimental evaluation of an ongoing job training program (see Orr et al. 1995). However, interest in quasi-experimental approaches remained.

Our knowledge of the strengths and weaknesses of quasi-experimental methods and ways of improving upon them has been advanced considerably by the availability of both control and comparison groups for two important programs, the National Supported Work Demonstration (NSW) that operated in the 1970s and the National JTPA Study. Researchers have been able to construct a series of after-the-fact comparison groups for NSW participants from the Panel Survey of Income Dynamics (PSID) and the Current Population Survey (CPS), while the National JTPA Study design featured an Eligible Non-Participant (ENP) sample for the explicit purpose of studying the efficacy of quasi-experimental methods. In both cases, the availability of experimentally based control groups and resulting net impact estimates provided the necessary "benchmark" against which quasi-experimental approaches could be gauged.

Researchers analyzed various subsamples of the NSW data in the late 1980s. LaLonde (1986) and Fraker and Maynard (1987) concluded that, while replication of experimental results was better for some groups (i.e., females and AFDC recipients) than others (e.g., males) and when using longitudinal rather than cross-sectional data, quasi-experimental methods generally failed at replication even when econometric estimates passed conventional specification tests. Fraker and Maynard found that such estimates were sensitive to both the method of comparison group construction and the analytic model specified. They further stressed the importance of utilizing prior earnings and local labor market controls in quasi-experimental estimation. Heckman and Hotz (1989) optimistically offered relatively simple specification tests that would allow evaluators to choose among alternative comparison groups by testing whether program participation variables predicted *pre*-program earnings. Subsequent analyses by Heckman and others led them to advise the use of such tests only with considerable caution.

Dehejia and Wahba's (1999, 2002) analysis of NSW data suggested that evaluators could achieve relatively low bias from propensity score matching in job training programs. However, Smith and Todd's (2003) reanalysis suggests otherwise. Heckman, Ichimura and Todd (1997), Heckman, Ichimura, Smith and Todd (1996, 1998) have extensively analyzed the JTPA data and more recently Smith and Todd (2003) have reanalyzed the NSW data. This work is augmented by analysis of experimental and quasi-experimental data on the School Dropout Demonstration Assistance Program (SDDAP) by Agodini and Dynarski (2001). Among the key conclusions from research on the development of low-bias matching estimators are the following (e.g., Smith and Todd, 2003, pp. 2ff.):

- Evaluation data should include a rich set of variables on program participation and labor market outcomes;
- The quasi-experimental comparison group should be drawn from the same local labor market as the program participants;
- The dependent variable (e.g., employment, earnings) should be measured in the same way for both participants and non-participants;
- Specification tests such as those suggested earlier by Heckman and Hotz may be necessary but they are not likely to be sufficient to choose among alternative matching estimators in all cases;
- Difference-in-difference matching estimators perform substantially better than cross-sectional ones; and
- The choice of matching procedures (e.g., between nearest neighbor and local linear matching) generally does not have a strong or consistent effect on estimated biases.

Smith and Todd argue that evaluators have been mistakenly trying to answer the question which estimator is the best one always and everywhere, when they should be striving "to develop a mapping from the characteristics of the data and institutions available in particular evaluation contexts to the optimal quasi-experimental estimators for those contexts." We clearly lack the option of benchmarking our quasi-experimental methods to experimental ones to ensure that they are the best attainable. The WIA program has not been the object of an experimentally based evaluation to date, nor is it likely to be in the near future. We can, however, incorporate the more important lessons learned from preceding (and ongoing) research on quasi-experimental methods into our evaluation approach. Furthermore, we can build upon our understanding of and, in time, make fuller use of data on program participation in our participating states.

Approach

This section describes the approach used to produce the net impact estimates. The "treatment" in this study is receipt of training services by individuals who exited from WIA in Program Year 2000 (July 2000–June 2001). The counterfactual that we are using to construct a comparison group is that if there were no training services, then individuals would receive core and/or intensive services only. Thus the pools of observations from which we construct the comparison groups are comprised of individuals who exited from WIA in the same program year but did not receive training services.[5] A major issue is the extent to which participants who receive core or intensive services are systematically different (in observable and unobservable ways) from participants who receive training services.[6] Systematic differences between these two groups weaken the design of the study by threatening the assumption that participants who receive core or intensive services only are an appropriate pool from which to draw comparison observations. Our econometric techniques control for observable differences; however, there is little we can do to control for unobservables.

In one of the states, in addition to using training as the treatment, we have used a different "treatment" and a different counterfactual. In particular, we have used "WIA services" as the treatment and assumed that the alternative to WIA services was mediated or core services through Wagner-Peyser funded activities of the Employment Service. The net impact estimates from that state are presented in this report along with those from the other states where we used a "WIA training" treatment group.

<u>Comparison and Treatment Samples</u>. The key methodological issue that needed to be resolved was to determine the observations in the WIASRD data system who received core or intensive services that were most comparable to those who received training services through WIA. Let *T* (for treatment) denote the administrative data for the individuals who terminated from training services during PY 2000. Let *U* (for universe) denote the set of observations from

[5] All of our analyses have been done separately for Adult and Dislocated Worker participants.
[6] After all, the services are intended to be sequenced, so that the individuals receiving training services should be those having the most labor market difficulty *in that service area*. There does seem to be significant variation in who receives training services *across* areas, however.

which we chose the subset *C* (for matched comparison group). In the analysis for which we used data from the Employment Service, the ensuing methodological discussion holds, except that U and its subset, C, refer to ES microdata and T refers to the entire WIASRD sample for that state.

The basic idea is to have *C* be comprised of the observations where the individuals are most "like" the individuals comprising *T*. Fortunately, because all of the data come from the WIASRD system, there is substantial overlap in the variables that are in the data sets, such as age, race/ethnicity, education at program entry, disability status, ESL status, gender, region of state, veteran status, and prior employment and earnings history. The actual dimensions used for matching were derived from a number of WIASRD data elements as well as UI wage histories in the quarters preceding entry into the WIA program. The matching dimensions are listed in Appendix Table A.1.

Table 1 provides sample size information from the seven states. The first two columns of numbers show adult cases in the comparison group (core or intensive services) and in the treatment group (training services). The latter two columns provide the same information for dislocated worker cases. Note that these sample sizes do not reflect exclusions for such reasons as the absence of wage record data or missing data items.

At the state level, there are substantial sample sizes and a reasonably good balance between the size of the treatment population and the comparison pool. However, in order to draw the comparison sample from the same local labor market as the treatment group, we needed to investigate the sample size and the balance between comparison and treatment group at *each* of the local workforce development areas. In some of the local areas, there was not an adequate sample to proceed with the methodology. Two problems arose. Either there was a very small number of cases in the treatment group or in the comparison group pool, or there were not enough cases in the comparison group relative to the treatment group so that there would be reasonably "good" matches for the treatment group observations. In the former case, we combined contiguous local areas to maintain as closely as possible the local labor market area. Our exact combinations are available on request, but the result was to aggregate the 24 local areas in Florida to 11 (of which 6 are the original single area), aggregate 20 local offices in Georgia to six (of which one is a single area), aggregate the 26 local areas in Illinois to 8 (of which one is a single area), aggregate the 11 areas in Maryland to 6 (of which 4 are single areas), aggregate the 14 areas in Missouri to 13 (of which 12 are single areas), aggregate the 28 areas in Texas to 21 (of which 14 are single areas), and to aggregate the 12 areas in Washington to 9 (of which 6 are single areas).

Table 1

WIASRD Cases by Type of Services Received, by State

State	Adult, core or intensive	Adult, training	Dislocated, core or intensive	Dislocated, training
Florida	5,274	5,348	2,009	4,022
Georgia	521	700	196	698
Illinois	1,103	1,589	1,281	2,160
Maryland	1,492	630	705	1,141
Missouri	2,398	869	1,302	779
Texas	3,487	5,855	3,566	2,474
Washington	401	845	437	1,329

After aggregating the local workforce areas, we had adequate overall sample sizes by local area, but in some cases the balance between treatment and comparison group was not adequate. We (arbitrarily) set a condition of having at least twice as many observations in the comparison group pool as in the training (treatment) group. If this condition was not met, we drew a random sample from the treatment group. Specifically, we chose a "sampled treatment" population that was one-half as large as the comparison pool. Or, in other words, on average each treatment case would have two comparison cases from which to match. After aggregation of the local areas, there were a total of 74 areas across the seven states. We drew a sample of treatment cases in 60 areas for adults and 62 areas for dislocated workers.[7]

Statistical matching. Following the literature, we have chosen to perform the statistical matching of observations using "propensity score" matching with replacement.[8] The propensity

[7] Note that we kept a weight whenever we sampled, so that results could be presented on a weighted basis. This document presents unweighted results.

[8] We also experimented with a matching technique called weighted multivariate matching. Weighted multivariate

score approach uses the standard technique of pooling T and U, and estimating the probability of being in T using logistic regression (see Dehejia and Wahba 1998). For each observation, the predicted probability of being in the treatment group is called a propensity score, and treatment observations are matched to observations in the comparison sample with the closest propensity scores. The software program STATA was used for the propensity score matching.

If there are systematic differences between the observations in T and U, and if the model is reasonably good, then the predicted propensity scores for the treatment observations will tend toward unity, and the predicted values in the comparison group pool will tend toward zero. Matching occurs by considering each treatment observation and minimizing (over all observations in U) the difference between its propensity score and the propensity score of the observations in U. In practice, we found a few matched pairs where the difference in the predicted propensity score was relatively large, and so we also implemented "caliper matching," deleting from consideration any treatment observation for which the match difference exceeded the chosen "caliper" level.[9]

Table 2 provides information regarding the quality of the participation logit model. For each program (adults and dislocated workers), the table displays three pieces of information: the "naïve" predicted propensity, the mean predicted propensities for the treatment group and the matched comparison group, and a statistic we refer to as the 20th percentile indicator. The "naïve" prediction is simply the share of the estimation population that belongs to the treatment group. It represents the share of the time we would get it correct if we predicted training participation totally randomly. This "naïve" predictor is a benchmark against which to compare

---

matching considers all available information in the treatment and comparison pools (as opposed to a functional combination of information) and places greater importance on those dimensions for which the treatment group differs systematically from the group of potential comparators. It essentially uses multivariate distance to judge similarity, but before doing so, it uses regression coefficients from a propensity score estimation to expand or shrink the multidimensional space in numerous ways, so that greater emphasis is placed on those dimensions that make the greatest unique contributions to the pre-existing differences between groups. Distance is defined using the Mahalanobis metric across all of the dimensions of matching variables, but unlike multivariate nearest neighbor algorithms, weighted multivariate matching weights each of the components distances in the Mahalanobis metric by its absolute standardized regression coefficient (from the logistic regression predicting treatment) before summing.

[9] The calipers that we chose for the propensity score matching differed by state and by program type. Specifically, we used min (90th percentile, 0.01), where 90th percentile is for the distribution of pairwise differences in propensities. In other words, we never allowed the predicted propensities to differ by more than 1 percentage point. In most states, the caliper was less than .01. The caliper matching eliminated roughly 10 percent of the matched sample. The caliper for the weighted multivariate method was set at the 90th percentile of distances for all subsamples.

the mean predicted propensities using the model. The predicted mean for the comparison group should be less than the naïve predictor, and the predicted mean for the treatment group should be greater. The 20th percentile indicator represents the percentile of the comparison group sample for the predicted propensity at the 20th percentile of the treatment predicted propensity score distribution. Battelle (1997) hypothesizes that a value of 80.0 percent represents a "good" participation model.

Table 2

Indicators of Participation Model Quality

State	Adult				Dislocated Worker			
	Naïve, mean	Comp. sample, mean	Treat. sample, mean	20th %ile indic.	Naïve, mean	Comp. sample, mean	Treat. sample, mean	20th %ile indic.
State 1	0.312	0.264	0.416	42.8%	0.316	0.279	0.395	38.7%
State 2	0.333	0.312	0.374	39.7%	0.310	0.285	0.365	44.6%
State 3	0.165	0.139	0.298	61.8%	0.257	0.211	0.389	53.8%
State 4	0.192	0.126	0.472	84.2%	0.314	0.267	0.417	50.1%
State 5	0.300	0.254	0.408	54.6%	0.318	0.275	0.410	47.0%
State 6	0.377	0.291	0.517	57.7%	0.277	0.141	0.631	86.3%
State 7	0.289	0.243	0.401	56.6%	0.266	0.225	0.379	58.2%

The overall picture suggested by the statistics in Table 2 indicates that the observed explanatory variables are not strong at discriminating between who gets training and who does not. The average predictions, for the most part, are not all that better than the "naïve" model, and the 20th percentile indicator is rarely over 60 percent.

Another method for ascertaining the quality of the matches is to test more directly for systematic differences between the observations in T and U by means of a logistic regression predicting treatment from the matching dimensions. This method also carries the advantage of allowing a direct comparison between the quality of matches produced by the propensity score and weighted multivariate methods. Results of these tests are presented in Table 3, summarized in the form of a multiple R-squared statistic for each model. Note that a higher R-squared indicates greater differences between T and U on the baseline dimensions.

From this table it is clear that both matching methods substantially reduce the differences between the treatment and matched comparison samples. Neither matching method appears to be consistently more effective than the other in match quality produced.[10] Nor does the reduction in differences between treated and matched samples appear to be consistent across states. However, there is evidence of a pattern that may begin to explain why one method sometimes produces better quality matches than the other. In general, those cells with larger sample sizes have the best quality matches when using the propensity score method, while the relatively smaller samples (though still quite large by most standards) fare better using the weighted multivariate technique. This can be seen, for example, using the sample sizes from Table 4b (below). Those cells for which weighted multivariate matching produced smaller R-square values in Table 3 have an average of just over 500 participants, while those for which propensity scoring yielded better matches averaged more than 1,100 participants. This pattern, if it were to hold in future research, would suggest one answer to Smith and Todd's (2003) recommendation that quasi-experimental analysis techniques be mapped to those evaluation situations for which they are best suited.

[10] Note that several differences in the implementation of the two matching methods may account for some of the differences in match quality. Importantly, the weighted multivariate method used sampling *without* replacement, and did not include employer industry dummy variables among the baseline matching dimensions.

Table 3

Comparing Match Quality of Propensity Score and Weighted Multivariate Methods

State	Adults unmatched sample	Adults Propensity Score match	Adults Weighted Multivariate match	Dislocated workers unmatched sample	Dislocated workers Propensity Score match	Dislocated workers Weighted Multivariate match
State 1	12.7%	9.3%	6.4%	9.1%	7.6%	3.1%
State 2	5.1%	0.5%	1.0%	6.7%	0.4%	1.4%
State 3	16.7%	1.6%	2.5%	15.3%	3.0%	4.8%
State 4	33.9%	5.2%	8.3%	11.9%	4.0%	3.0%
State 5	13.1%	2.0%	2.7%	11.1%	0.7%	4.2%
State 6	18.2%	3.6%		44.5%		
State 7	13.8%	0.3%	1.6%	13.3%	2.1%	0.7%
State 1/ES	29.5%	1.1%	2.1%	18.7%	0.4%	0.8%

Entries are pseudo R-squareds in percentage terms.

Net impact estimators. After constructing the T and C matched samples, we estimated the net impacts of the program using several different estimators. There are many potential estimators that have different properties and that make different assumptions about the data. Let us start the discussion with simple (unconditional) differences in outcome means. This nonparametric approach suggests that the net impact can be fully estimated by averaging the differences in outcomes for each individual matched pair of observations in T and C. Suppose that average quarterly earnings is one of the outcome variables of interest. Then the net impact per participant would be estimated as follows:

(1) $$Y = \sum_j \{(ET_j - EC_j)\} / n$$

where ET_j = the average quarterly earnings (adjusted to constant \$) after exiting intensive or training services for the jth individual

EC_j = the average quarterly earnings (adjusted to constant \$) after the appropriate program year for the individual(s) in the comparison group matched to j

n = the number of individuals in the sampled treatment and comparison groups

This estimate of the program's net impact assumes that the treatment sample and matched comparison group are not systematically different from each other. However, we know that the matching technique is only able to control for observables, and there may be unobservable differences between them. A second approach would be to difference out unobservables through a difference-in-differences technique. That is, we estimate the program's net impact as follows:

$$(2) \qquad Y = \sum_j \{(ETPOST_j - ETPRE_j) - (ECPOST_j - ECPRE_j)\} / n$$

where, *ECPREj*, *ETPREj* = average quarterly earnings of the *j*th individual (and his/her match) prior to being served

ECPOSTj, *ETPOSTj* = average quarterly earnings of the *j*th individual (and his/her match) after participating

This is again a nonparametric approach that makes few assumptions about the data or earnings mechanisms.

The difference-in-means approaches in equations (1) and (2) assume that the matching technique has eliminated differences between the sampled treatment and matched comparison cases that may affect program outcomes. This is a strong assumption to make with a quasi-experimental approach to net impact evaluation. Consequently, we have estimated regression-adjusted differences in means. This parametric approach assumes that we can use observed variables to control for differences between the matched pairs of observations. A very simple model for regression-adjustment is displayed in the following equation:

$$(3) \qquad Y_j = a + B\text{N}\, X_j + c\, T_j + u_j$$

where, Y_j = outcome for individual *j* (or for the individual(s) matched to individual *j*); that is $ETPOST_j$, or $ECPOST_j$

X_j = vector of variables describing individual *j* (or for the individual(s) matched to individual *j*) that are thought to be correlated to the outcome Y_j

T_j = 1 if individual *j* is in the participant sample and 0 if not

u_j = error term, usually assumed to have a mean of 0 and standard deviation of 1.

The parameter estimate c would be the net impact of participation in the program. Note that we have also estimated (3) with pre- and post- changes in the outcomes as dependent variables.

That is, the Y_j equal ($ETPOST_j - ETPRE_j$) or ($ECPOST_j - ECPRE_j$) depending on whether the observation is in the sampled treatment or matched comparison group.

The regression-adjusted model in equation (3) needs to be examined very carefully to understand the underlying assumptions. First of all, most of the X_j were used in constructing the matched sample. We have assumed in those cases that the X variables are important in determining program eligibility (matching process) and in determining program outcomes (equation (3)). If all of the X variables were used in the matching process and in the outcome equations, and if all of the matches were "perfect," e.g. a 29-year-old male with seven years of work experience in a specific industry is matched with a 29-year-old male with seven years of work experience in the same industry, then c will be equal to the net program impact estimator in (1). As mentioned, the reason to estimate (3) is because the quasi-experimental approach is likely to have "imperfect" matches.

<u>Preferred specification.</u> The preferred specification for the adult population, and is used in the tables below, is equation (3), with difference-in-differences as the dependent variable. Because dislocated workers, by definition, have a discontinuity in their labor force characteristics, the difference-in-differences approach is less valid, and so the preferred specification for this population is equation (3) using outcome levels.

Results

Outcomes. Our project will ultimately estimate net impacts for labor market as well as other outcomes of interest. This report presents results for the former, i.e., labor market outcomes. Specifically, we have estimated earnings and employment impacts as measured by UI wage records over available post-WIA quarters of data (i.e., through 2002Q2) following exit. Note that this means we have between four and seven post-WIA quarters of data for all states.

The four *labor market-related outcomes* examined in this study are:

- Employment in the fourth quarter after exit**,** defined as having over $100 in total earnings;
- Fraction of employment during post-exit quarters two – four, defined as 0, .33, .67, or 1.0 if the observation has total earnings over $100 in zero, one, two, or three of the

relevant quarters;

- Earnings in the fourth quarter after exit, provided they were over $100; and
- Average quarterly earnings for the second, third, and fourth quarter after exit conditional on total earnings over $100, i.e. if quarterly earnings are greater than $100 for only two of the three quarters, then the average is calculated for two quarters.

Difference-in-differences estimators were calculated for the last three outcomes using pre-program entry quarters 3 through 6 as the base period. The logic behind this choice is that pre-program quarters 1 and 2 may be subject to pre-program labor market "distress" (Ashenfelter's 1978 earnings "dip"), so that these four quarters (−3 through −6) may be the closest annual period to program entry for which pre-program circumstances can be calculated[11].

Table 4a provides initial net impact estimates for adults and dislocated workers on conditional earnings in the fourth quarter after exit using propensity score matching and the preferred specifications (regression-adjusted mean difference-in-differences for adults and regression-adjusted mean comparisons for dislocated workers).

[11] For the results in Table 4b, the difference-in-difference estimator was not used. In lieu of this, the roughly equivalent approach of using base period wages as a covariate was used.

Table 4a

Net Impact[12] Estimates of WIA Training Services on Conditional Earnings in Quarter + 4, Propensity Score Matching

State	Net impact for adults	Sample size	Net impact for dislocated workers	Sample size
State 1	$ 621.9 (516.0)	97	--$505.0 (678.1)	214
State 2	$ 625.7*** (127.8)	1,590	$ 202.5 (156.9)	2,171
State 3	$ 340.9 (236.9)	396	--$629.1** (273.1)	509
State 4	$ 474.6 (306.4)	297	$ 229.0 (372.8)	316
State 5	--$ 9.3 (312.2)	384	$ 430.1* (258.7)	810
State 6	$ 576.2* (302.6)	133	-----	-----
State 7	$ 288.1** (130.9)	1,624	$ 150.1 (195.6)	877

Conventional OLS standard error in parentheses. ----- sample size too small to generate valid estimates.

Table 4b presents estimates of the impact of WIA training on adults and dislocated workers, but this time using the weighted multivariate matching method to select members of the comparison group.

[12] Regressors for statistical adjustment of these and other impacts are listed in Appendix Table A.1.

Table 4b

Net Impact Estimates of WIA Training Services on Conditional Earnings in Quarter + 4, Weighted Multivariate Matching

State	Net impact for adults	Sample size	Net impact for dislocated workers	Sample size
State 1	$807**	203	$283	455
	(392)		(293)	
State 2	$852***	2002	$164	2251
	(116)		(154)	
State 3	$337*	554	-$109	513
	(192)		(249)	
State 4	-$369	427	$194	425
	(257)		(345)	
State 5	$432*	593	$102	752
	(225)		(246)	
State 6	n/a		n/a	
State 7	$504***	1817	$42	955
	(115)		(178)	

Conventional OLS standard error in parentheses. State 6 data not yet released to this research team.

Table 5a provides net impact estimates for employment in the 4^{th} quarter after exit, when measured using the propensity score matched comparison group. Here there are no difference-in-difference estimators. Furthermore, because the dependent variable is either 0 or 1, the regression adjustment is done with logit. Note that the net impacts are measured in percentage-point changes, not percentages.

Table 5a

Preliminary Net Impact Estimates of WIA Training Services on Employment in Quarter + 4, Propensity Score Matching

State	Net impact for adults	Sample size	Net impact for dislocated workers	Sample size
State 1	11.9%	182	-- 2.6 %	278
State 2	4.7 %**	2,466	-- 2.4 %	2,766
State 3	1.1 %	678	-- 6.1 %*	636
State 4	-- 1.1%	416	-- 2.2 %	448
State 5	-- 1.4 %	652	1.8 %	1,028
State 6	--20.6%	210	-----	
State 7	-- 1.9%	2,910	1.1%	1,206

*** significant at .01 level; ** significant at .05 level; * significant at .10 level (Chi-square test of significance on logit coefficient). ---- sample size too small to generate valid estimates.

Table 5b displays net impacts on employment at four quarters after exit, when calculated using the weighted multivariate matched comparison group.

Table 5b

Preliminary Net Impact Estimates of WIA Training Services on Employment in Quarter + 4, Weighted Multivariate Matching

State	Net impact for adults	Sample size	Net impact for dislocated workers	Sample size
State 1	2.1%	340	-0.5%	684
State 2	4.3%**	2881	-2.2%	2898
State 3	-0.6%	820	-3.7%	664
State 4	-3.4%	573	-6.4%*	580
State 5	-5.2%	958	1.8%	956
State 6	n/a		n/a	
State 7	-0.3%	2994	-3.5%	1316

*** significant at .01 level; ** significant at .05 level; * significant at .10 level (Chi-square test of significance on logit coefficient). State 6 data not yet released to this research team.

Net Impacts of WIA as Treatment. As noted, for one of the states, we were able to access the Employment Service administrative data, which allowed us to treat the entire WIA services sample as the "treatment" group. The net impact estimates for this exercise are provided in Tables 6a and 6b.

Table 6a

Preliminary Net Impact Estimates of WIA (Core, Intensive, or Training) Services on Employment and Conditional Earnings in Quarter +4, Propensity Score Matching

State	Net earnings impact for adults	Sample size	Net earnings impact for dislocated workers	Sample size
State 1	$736.0*** (194.7)	882	$312.3** (151.9)	4,048
	Net employment impact for adults	Sample size	Net employment impact for dislocated workers	Sample size
State 1	5.9%**	1,739	10.8%***	2,912

Conventional OLS standard error in parentheses.

Table 6b

Preliminary Net Impact Estimates of WIA (Core, Intensive, or Training) Services on Employment and Conditional Earnings in Quarter +4, Weighted Multivariate Matching

State	Net earnings impact for adults	Sample size	Net earnings impact for dislocated workers	Sample size
State 1	-$44 (167)	1097	$340* (182)	1249
	Net employment impact for adults	Sample size	Net employment impact for dislocated workers	Sample size
State 1	4.9%**	1947	13.6%***	1804

Conventional OLS standard error in parentheses.

Discussion

The net impact estimates show considerable variation across participant subgroup (i.e., adults and dislocated workers) and across states. However, across matching methods the results are remarkably consistent. The estimates in Tables 4a and 4b suggest that the treatment has a positive impact on quarterly earnings for adults. The estimates in the first columns of those tables suggest that the earnings impacts are between $300 to $600 (with the exception of state 5), which would be approximately 7 to 10 percent. However, it should be noted that only four or five of the seven estimates are significant. The earnings impacts for dislocated workers are quite mixed. Only two of the states had statistically significant impacts when using propensity score matching—one was positive and the other was negative. The weighted multivariate matching method showed no significant earnings impacts for dislocated workers.

The employment impacts for adults are also varied. Only one of the estimates is significant, and it is positive. Only one of the employment impacts for dislocated workers is significant, and it is negative. Both matching methods show the same pattern of employment effects for adults and dislocated workers, and the point estimates for the one positive and one negative effect are nearly identical.

The key issue to understand in analyzing the results is how local agencies allocate training services.

Administrative Records Opportunities and Limitations

There are both opportunities and limitations associated with the use of administrative records for estimating net impacts from WIA participation. Note that in discussing this issue, we are contrasting the use of an *array of linked administrative records* to alternatives such as longitudinal surveys of participants and either comparison or control group members. A number of groups, panels, and committees have examined this issue over the past decade or so, among them the Northeast-Midwest Institute (e.g., Bishop 1989, and King 1989), the National Council for Employment Policy (Baj, Trott, and Stevens 1991), the Advisory Panel on Research Uses of Administrative Data (1998), and the National Research Council (Moffitt and Ver Ploeg 2001). The good news is that there has been broad consensus on the advantages and disadvantages of using administrative data for research and evaluation purposes and seldom any real disagreement. Here we summarize the opportunities and limitations of the use of administrative data for impact estimation based on this literature as well as our own experience in the ADARE project.

Major advantages to using administrative records for impact estimation include the following:

- *Cost.* Administrative data such as WIA, ES, and UI wage records can be tapped at a small fraction of the cost of generating similar information through surveys. For example, the cost of ongoing UI wage records collection was estimated to be as little as one-sixth the cost of securing employment and earnings data via participant surveys (Baj, Trott, and Stevens 1991, pp. 43-46).

- *Coverage.* Administrative records' coverage of participant demographics, services and activities, and outcomes tends to be both broad and deep when linked across multiple data sources. Some research and evaluation projects have analyzed linked data for as many as a dozen different federal/state programs.

- *Sample Sizes.* One of the outstanding features of administrative records is that they generally offer researchers the opportunity to conduct their analysis on the entire population of interest rather than simply restricted samples. Researchers can thus avoid 'thin' cell sizes when analyzing population subgroups, service strategies, or other groupings. Moreover, two of the more vexing problems for surveys are largely avoided: non-response bias and sample attrition over time.

- *Longitudinal Span.* Tapping UI wage records for employment and earnings information also allows researchers to track key labor market variables of interest before, during, and after an intervention with relative ease. Access to pre-program information is critical for quasi-experimental impact estimation in particular because it gives researchers the ability to match on extensive prior work history and to test for the presence of selection bias. Ashenfelter's earnings 'dip' can be modeled with such data, as can periods of unemployment prior to program entry. Long-term labor market outcome data—sometimes as long as 7-9 years—allow researchers to document any decay rates of earnings impacts as well as the potentially greater value of human capital development over labor force attachment approaches that might not be apparent using only short-term data.

- *Flexibility.* Administrative data also provides researchers with considerable flexibility to make mid-course modifications to their evaluation designs without having to start over or to model such important elements as training program participation decisions.

- *Increasing Access*. In earlier decades, it could be quite difficult for researchers to gain access to administrative records for participants, potential comparison or control group members and their labor market and related experiences. However, this began to change over the 1980s and 1990s as more researchers—including members of the ADARE project, as well as many others—began to develop ongoing partnerships with state agencies for data access that featured carefully structured agreements with confidentiality protections. Some of the ADARE research partners serve as data archivists for their states, while others only access data on an as needed basis.

- *Improving Quality.* As researchers have made increased use of administrative data, ranging from education and training and welfare to employment and earnings records, the quality of these data has improved. Attention and use tends to beget quality in most, if not all, cases. The fact that welfare is now time-limited and performance standards are being applied across multiple funding streams is reinforcing quality, with some important reservations that are noted below.

- *Accuracy/Consistency.* Administrative data tend to be more accurate and consistent than participant survey data for several reasons, including lack of recall and self-reporting bias problems. For example, UI wage records, unlike survey data, are auditable and subject to challenges by employers, on the one hand, who seek to minimize reported wages, and by employees, on the other, who seek to maximize them in support of their UI claim.

Important limitations of administrative data for impact estimation and evaluation include the following:

- *Costs of Conversion from Management to Research Data.* By definition, administrative records have been generated to support program administration and management, not research and evaluation. Converting such data, often linked across many different sources, can be costly and time consuming. Data must be cleaned and researchable data sets created with care to document definitions, missing data, and other features.

- *State Capacity for and Expertise in Data Sharing.* States still vary widely in terms of their capacity to negotiate data sharing agreements and provide data access to researchers. This situation has improved a great deal over the past decade, but state capacity remains a limitation.

- *Limited Content.* Another limitation relates to limited content in some administrative data sets. Data elements that might be highly desirable from a research and evaluation vantage point may be lacking altogether. For example, many state and local education agencies allow students to use unique numbers rather than Social Security Numbers (SSNs) as student identifiers, precluding linkages to other data sets (e.g., UI wage records). And, while powerful, UI wage records in nearly all states lack data elements for hourly wage rates, hours, and occupation, all of which would be very helpful for evaluation purposes.

- *Incomplete Archiving.* While some states began archiving their program and work history data many years ago—some in the early 1980s (e.g., Illinois, Missouri) or late 1980s (e.g., Texas)—most did not. The absence of complete archival records may preclude high-quality impact estimation with matching on prior work history and requisite testing for selection bias in many states.

- *Coverage.* Coverage tends to be quite good in most administrative records, but it is far from ideal. For example, WIA's somewhat loose reporting provisions, flexible registration and exit points, and high performance expectations, have provided incentives for under-reporting participation and have limited its coverage of self-service activities. Similarly, not all employment is covered by UI, creating coverage gaps. Key gaps include religious, railroad, and other areas.

- *Inter-item Quality Variation.* Data items required for program eligibility or other reasons tend to be of much higher quality than those that are not. Statistical problems will arise to the extent that lower quality data elements (e.g., educational attainment or other demographics) are important covariates.

- *Reporting Unit.* In general, the reporting unit for most administrative data is the individual participant, student, or employee. Some analyses, on the other hand, are most appropriately focused on the household or family. Linking individual-level data to aggregate to households or families may be difficult.

- *Confidentiality and SSNs.* Issues surrounding privacy and confidentiality, especially regarding students and educational interventions, have become seemingly more intractable recently, despite rulings by the U.S. Supreme Court in mid-2002 that would appear to remove some of the potential liability at least at the postsecondary level. Recent (January 2003) guidance from the U.S. Department of Education is, unfortunately, having the effect of discouraging record sharing across funding streams.

- *Conflicting Time Periods/Frequency.* The problem of conflicting time periods or frequency in various administrative records sources is a continuing one, but not one that is usually fatal for analysis. Typically, it is possible to aggregate monthly data (e.g., welfare or WIA) to quarterly in order to utilize quarterly UI wage records for measuring outcomes or estimating employment or earnings impacts.

- *Inter-state and Inter-area Comparability.* Inter-state and inter-area data comparability is an important concern for researchers and evaluators, especially as federal policy devolution has picked up speed in the 1990s and early 2000s (see Barnow and King, 2003). WIA presents a case in point with variations in definitions, data collection points and procedures, data elements, and reporting practices, both among and within states. Such variations can adversely affect the precision of any resulting impact estimates and, at the least, may constrain the types of analyses that are possible to the lowest common denominator.

- *Inter-state UI Wage Records Data Access.* Although the Wage Records Interchange System (WRIS) project operated by the National Association of Workforce Agencies represents a commendable effort to address the issue of inter-state UI wage records access, it is far too limited in scope and scale at present. Too few states are volunteering to participate in WRIS, and it remains a WIA-focused endeavor, seriously limiting its potential utility.

- *Rise of Internet-based and Self-directed Services.* Jobseekers and employers increasingly rely on the Internet and self-directed services, leaving large gaps in our knowledge base about workforce and related services and their impacts. States such as Utah are experimenting with mechanisms for capturing more of this activity, but most are only worrying about it.

Despite their flaws, linked administrative records, on balance, offer researchers attractive opportunities for estimating the impact of WIA and related services at relatively low cost. Administrators at all levels should strive to improve the quality and accessibility of these data in the near future, while ensuring that appropriate protections are in place to ensure individual privacy and confidentiality. Areas that rank high on our own short list of desired improvements include the following, among others:

- Tighter, more comprehensive reporting under WIA, including the addition of applicant data;
- Resolution of the FERPA issues at the federal level, hopefully so that analysts can have improved access to postsecondary student records for research and evaluation purposes;
- Expansion of WRIS to encompass all states and to include coverage of most major federal workforce and education funding streams; and
- Federally supported technical assistance and training to bolster state capacity for data sharing and access for research and evaluation.

Appendix Table A.1

Baseline Dimensions[13] for Matching and Regressors for Statistical Adjustment of Impacts

Dimension	Description and rationale
Local board code	As reported in WIASRD item 301. Exact matches are required on this dimension (or an aggregated version).
Time of exit	Quarter of WIA exit date, item 303.
Age at entry, years	Based on birth date (item 102) and WIA registration date (item 302).
Gender	Binary based on item 103: 1=female, 0=male.
Disability	Binary based on item 104: 1=yes (any), 0=no.
Ethnicity Hispanic	Binary based on item 105: 1=yes, 0=no.
Eth. Black or African American	Binary based on item 108: 1=yes, 0=no.
Eth. White	(omitted category). Binary based on item 110: 1=yes, 0=no.
Veteran	Binary based on item 111: 1=yes (any), 0=no.
Employed at registration	Binary based on item 115: 1=yes, 0=no.
Limited English	Binary based on item 116: 1=yes, 0=no.
Single parent	Binary based on item 117: 1=yes, 0=no.
Unemployment compensation claimant or exhaustee	Binary based on item 118: 1=yes, 0=no.
Low income	Binary based on item 119: 1=yes, 0=no.
TANF, GA, RCA, or SSI recipient	Binary based on items 120 and 121: 1=yes (either), 0=neither. Some states do not have G.A.
Education less than high school	Education binary based on item 123, highest grade completed. 1=less than high school graduate or GED, 0=greater.
Education high school graduate	(omitted category). Education binary based on item 123. 1= high school graduate or GED, 0=lesser or greater
Education beyond high school	Education binary based on item 123. 1=some college or greater, 0=less.
Employment	Percent of quarters employed in pre-WIA quarters 3-8, beginning with first employment in pre-WIA interval
Conditional earnings	Average earnings in pre-WIA quarters 3-8, of those quarters in which employed.
Earnings trend	Linear trend in earnings in pre-WIA quarters 3-8.
Earnings variation	Coefficient of variation of earnings in pre-WIA quarters 3-8.

[13] Due to data limitations, not all dimensions were used for every combination of state and subpopulation (adults or dislocated workers), nor were all dimensions used for both matching methods.

Dimension	Description and rationale
Turnover	Average number of employers per quarter in pre-WIA quarters 3-8.
Earnings dip	Change in average quarterly earnings (unconditional) from pre-WIA quarters 3-8 to second and first quarters prior to entry.
Employer industry	Industry code for most recent employer as of WIA entry, parameterized with a small number of dummy variables.

References

Agodini, Roberto and Mark Dynarski (2001). “Are Experiments the Only Option? A Look at Dropout Prevention Programs,” Princeton, New Jersey: Mathematica Policy Research, Inc., August.

Ashenfelter, Orley (1978). “Estimating the Effect of Training Programs on Earnings,” *Review of Economics and Statistics* 60(1), pp. 47-57.

Baj, John, Charles E. Trott, and David W. Stevens (1991). *A Feasibility Study of the Use of Unemployment Insurance Wage-Record Data as an Evaluation Tool for JTPA,* Washington, D.C.: National Commission for Employment Policy, Research Report Number 90-02, January.

Barnow, Burt S. (1987). "The Impact of CETA Programs on Earnings: A Review of the Literature," *The Journal of Human Resources*, Volume XXII, Number 2 (Spring), pp. 157-193.

Battelle Memorial Institute (1997). "Net Impact Evaluation: Appendix A: Technical Appendix," Unpublished manuscript.

Bishop, John (1989). “Policy Evaluation with Archived Wage Record Data: Limitations of Existing Data Sets,” In Northeast-Midwest Institute, *The Feasibility of a National Wage Record Database: Four Working Papers,* Washington, D.C.: The Northeast-Midwest Institute, The Center for Regional Policy, January 10.

Bowman, William R. (1993). *Evaluating JTPA Programs for Economically Disadvantaged Adults: A Case Study of Utah and General Findings,* Washington, D.C.: National Commission for Employment Policy Research Report 92-02. June.

Dehejia, Rajeev H. and Sadek Wahba (1999). “Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association,* Vol. 94, December, pp. 1053-.

Dehejia, Rajeev H. and Sadek Wahba (2002). “Propensity Score-Matching Methods for Nonexperimental Causal Studies,” *The Review of Economics and Statistics,* 84(1), February, pp. 151-161.

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

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

Friedlander, Daniel, David H. Greenberg, and Philip K. Robins (1997). “Evaluating Government Training Programs for the Economically Disadvantaged.” *Journal of Economic Literature* XXXV (December), pp. 1809-1855.

Heckman, James J. and V. Joseph Hotz (1989). "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs," *Journal of the American Statistical Association,* Vol. 84, pp. 862-874.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd (1998). "Characterizing Selection Bias Using Experimental Data," *Econometrica,* Vol. 66(5), September, pp. 1017-1098.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluation a Job Training Programme," *Review of Economic Studies,* Vol. 64, pp. 605-654.

Heckman, James J., Robert J. LaLonde and Jeffrey A. Smith (1999). "The Economics and Econometrics of Active Labor Market Programs," In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics,* Vol. 3, Amsterdam: Elsevier, pp. 1866-1932.

Hollenbeck, Kevin et al. (2001). *Net Impact and Cost-Benefit Evaluation of Washington State's Workforce Training System: A Proposal,* Kalamazoo, Michigan: W. E. Upjohn Institute for Employment Research.

Hollenbeck, Kevin et al. (2002). *Net Impact and Cost-Benefit Evaluation of Washington State's Workforce Training System: Final Report,* Kalamazoo, Michigan: W. E. Upjohn Institute for Employment Research.

Hotz, V. Joseph, Robert Goerge, Julie Balzekas, and Francis Margolin, Eds. (1998). *Administrative Data for Policy-Relevant Research: Assessment of Current Utility and Recommendations for Development, A Report of the Advisory Panel on Research Uses of Administrative Data,* Chicago, IL: Northwestern University/University of Chicago Joint Center for Poverty Research.

King, Christopher T. (1989). "Employment and Earnings Dynamics in the 1990s: Policy Issues and Potential Uses of Nationally-Archived UI Wage Records," In Northeast-Midwest Institute, *The Feasibility of a National Wage Record Database: Four Working Papers,* Washington, D.C.: The Northeast-Midwest Institute, The Center for Regional Policy, January 10.

King, Christopher T. et al. (1994). *Texas JOBS Program Evaluation: Final Report,* Austin: Center for the Study of Human Resources, Lyndon B. Johnson School of Public Affairs, The University of Texas at Austin.

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review,* Volume 74, Issue 4, September, pp. 604-620.

LaLonde, Robert J. (1995). "The Promise of Public Sector-Sponsored Training Programs," *Journal of Economic Perspectives* 9(2): 149-168.

Moffitt, Robert A. and Michele Ver Ploeg, Eds. (2001). *Evaluating Welfare Reform in an Era of Transition,* Washington, D.C.: National Academy Press, Report of the Panel on Data and Methods for Measuring the Effects of Changes in Social Welfare Programs, National Research Council.

Mueser, Peter, Troske, Kenneth R., and Gorislavsky, Alexey (2003). *Using State Administrative Data to Measure Program Performance,* draft, May 2003.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin, and George Cave. (1995). *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study,* Washington, D.C.: The Urban Institute Press.

Smith, Jeffrey and Petra Todd (2003). "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" Draft, May.

Stevens, David W. (2002a). *Mapping WIA One-Stop Client Flows: WIASRD Data Elements Needed,* Baltimore: The Jacob France Institute, University of Baltimore, July 16.

Stevens, David W. (2002b). *Workforce Investment Act Standardized Record Data (WIASRD) Issues for Evaluation Studies,* Baltimore: The Jacob France Institute, University of Baltimore, April 7.

Stromsdorfer, E., H. Bloom, R. Boruch, M. Borus, J. Gueron, A. Gustman, P. Rossi, F. Scheuren, M. Smith and F. Stafford (1985). *Recommendations of the Job Training Longitudinal Survey Research Advisory Committee,* Washington, D.C.: Employment and Training Administration, U.S. Department of Labor.