Empirical Economics (2007) 32:491–528
DOI 10.1007/s00181-006-0095-0

ORIGINAL PAPER

Evaluating multi-treatment programs: theory and evidence from the U.S. Job Training Partnership Act experiment

Miana Plesca · Jeffrey Smith

Accepted: 30 August 2006 / Published online: 19 April 2007

Abstract This paper considers the evaluation of programs that offer multiple treatments to their participants. Our theoretical discussion outlines the tradeoffs associated with evaluating the program as a whole versus separately evaluating the various individual treatments. Our empirical analysis considers the value of disaggregating multi-treatment programs using data from the U.S. National Job Training Partnership Act Study. This study includes both experimental data, which serve as a benchmark, and non-experimental data. The JTPA experiment divides the program into three treatment "streams" centered on different services. Unlike previous work that analyzes the program as a whole, we analyze the streams separately. Despite our relatively small sample sizes, our findings illustrate the potential for valuable insights into program operation and impact to get lost when aggregating treatments. In addition, we show that many of the lessons drawn from analyzing JTPA as a single treatment carry over to the individual treatment streams.

Keywords Program evaluation · Matching · Multi-treatment program · JTPA

An earlier version of this paper circulated under the title "Choosing among Alternative Non-Experimental Impact Estimators: The Case of Multi-Treatment Programs".

M. Plesca
Department of Economics, University of Guelph, Guelph, ON, Canada N1G 2W1
e-mail: miplesca@uoguelph.ca

J. Smith (✉)
Department of Economics, University of Michigan, 238 Lorch Hall,
611 Tappan Street, Ann Arbor, MI 48109-1220, USA
e-mail: econjeff@umich.edu

1 Introduction

In contrast to, say, clinical trials in medicine, many social programs, especially active labor market programs, embody heterogeneous treatments. Individuals who participate in such programs receive different treatments, at least in part by design. In this paper, we consider the implications of this treatment heterogeneity for program evaluation.

In our conceptual discussion, we examine the links between the level of treatment aggregation in an evaluation and the parameters of interest, the evaluation design, the available samples sizes (and therefore the precision of the resulting impact estimates) and the overall value of the knowledge gained from the evaluation. We raise the possibility of misleading cancellation arising from aggregating treatments with positive and negative (or just large and small) mean impacts.

We also present empirical evidence from an important evaluation of a multi-treatment program in the United States: the Job Training Partnership Act (JTPA). Our data come from an experimental evaluation denoted the National JTPA Study (NJS), which included the collection of "ideal" data on a non-experimental comparison group at some sites. Using the NJS data, we consider the impacts of disaggregated treatment types, and look for evidence of cancellation in the overall program impact estimates. As participants play an important role in determining treatment type in JTPA, we also look for differences by treatment type in the determinants of participation that might result from differences in the economics motivating participation. Finally, we examine the performance of non-experimental matching estimators applied to the three main treatment types in the JTPA program using the experimental data as a benchmark. Taken together, these analyses allow us to see the extent to which some of the lessons learned in related analyses that regard JTPA as a single aggregated treatment carry over to the disaggregated treatments. Our empirical analysis also adds (at the margin) to the literature on applied semi-parametric matching methods and, unfortunately, also illustrates the loss of precision that comes from disaggregating by treatment type. We find that many of the conclusions drawn from research that treats JTPA as a single treatment remain valid when looking at disaggregated treatments. At the same time, differences emerge when disaggregating that illustrate the value of doing so.

The remainder of the paper proceeds as follows. Section 2 provides a conceptual discussion of issues related to disaggregation by treatment type. Section 3 describes the evaluation design and the NJS data, while Sect. 4 describes the econometric methods we employ. Section 5 presents our empirical results and Sect. 6 concludes.

2 Treatment aggregation and program evaluation

Most active labor market policies include a variety of treatments. The JTPA program studied here offers classroom training in many different occupational

skills, subsidized on-the-job training at many different private firms, several types of job search assistance from various providers, adult basic education, subsidized work experience at various public or non-profit enterprises, and so on. Other countries also offer multiple service types to their unemployed. For example, in addition to the relatively standard fare offered by JTPA, Canada offers training in starting a small business, the New Deal for Young People (NDYP) in the United Kingdom offers participation in an "Environmental Task Force", the Swiss system studied in Gerfin and Lechner (2002) offers language training for immigrants, and Germany places some unemployed with temporary help agencies. In most countries, individuals get assigned to one of the multiple treatments via interaction with a caseworker, though some programs, such as the U.S. Worker Profiling and Reemployment Services System (WPRS) examined in Black et al. (2003), also make use of statistical treatment rules.

Moving from a program with one homogeneous treatment to a multiple treatment program greatly expands the set of possible questions of interest. In addition to the basic question of the labor market impacts of the program taken as a whole, researchers and policymakers will now want to know the impact of each treatment on those who receive it relative to no treatment and relative to other possible treatments. They will also want to know the effect of each treatment on those who do not receive it and they will likely want to know about and perhaps evaluate the system that allocates participants to treatments, as in Lechner and Smith (2007).

The existing literature applies non-experimental methods to answer all of these questions and experimental methods to answer some. Most experimental evaluations focus on estimating the impacts of treatments on those who receive them, though others, such as the U.S. Negative Income Tax experiments described in Pechman and Timpane (1975) and the Canadian Self-Sufficiency Project experiment described in Michalopolous et al. (2002), include random assignment to alternative treatments. The latter aids in answering questions regarding the impacts of treatments not actually received and the effect of alternative statistical treatment rules; see Manski (1996) for more discussion.

In thinking about evaluating the impact of treatments actually received on those who receive them, the key decision becomes how finely to disaggregate the treatments. Disaggregating into finer treatments avoids problems of cancellation in which the impacts of particularly effective treatments get drowned out by those of relatively ineffective treatments. At the same time, finer disaggregation implies either a loss of precision due to reduced samples sizes for each treatment or else a much more expensive evaluation (assuming reliance on survey data in addition to, or instead of, administrative records).

In practice, different evaluations resolve these issues differently. Consider the case of classroom training. Both the JTPA evaluation considered here and the evaluation of the NDYP in Dorset (2006) combine all classroom training into a single aggregate treatment. In contrast, the evaluation of Swiss active labor market policy in Gerfin and Lechner (2002) distinguishes among eight different services (five of them types of classroom training) along with

non-participation, and the evaluation of East German active labor market policy in Lechner et al. (2008) distinguishes among short training, long training and retraining (and non-participation). Perhaps not surprisingly, the German and Swiss evaluations both rely on administrative data, which allow much larger samples at a reasonable cost.

In experimental evaluations, choices about the level of disaggregation can interact with choices about the timing of randomization (and thereby with the cost of the experiment). In the NJS, the evaluation designers faced the choice of whether to conduct random assignment at intake, which occurred at a centralized location in each site, or at the many different service providers at each site. Random assignment at intake meant lower costs and less possibility for disruption, but it also meant assignment conditional on recommended services rather than on services actually initiated. As we document below, though clearly related, these differ substantially. Randomization later would have allowed the construction of separate experimental impacts for each provider (as well as various meaningful combinations of providers). In the end, cost concerns won out, with implications that we describe in Sect. 3.2.

3 Institutions, data and evaluation design

3.1 Institutions

From their inception in 1982 as a replacement for the Comprehensive Employment and Training Act to their replacement in 1998 by the Workforce Investment Act, the programs administered under the U.S. Job Training Partnership Act (JTPA) constituted the largest federal effort to increase the human capital of the disadvantaged. The primary services provided (without charge) under JTPA included classroom training in occupational skills (CT-OS), subsidized On-the-Job Training (OJT) at private firms and Job Search Assistance (JSA). Some participants (mainly youth) also received adult basic education designed to lead to a high school equivalency or subsidized "work experience" in the public or non-profit sectors. Eligibility for the JTPA program came automatically with receipt of means-tested transfers such as Food Stamps and Aid to Families with Dependent Children (AFDC — the main federal-state program for single parents) or its successor Temporary Aid to Needy Families (TANF). Individuals with family income below a certain cutoff in the preceding 6 months were also eligible for JTPA services (along with a few other small groups such as individuals with difficulty in English). The income cutoff was high enough to include individuals working full time at low wage jobs looking to upgrade their skills. Devine and Heckman (1996) provide a detailed description of the eligibility rules and an analysis of the eligible population they shaped.

As part of the "New Federalism" of the early Reagan years, JTPA combined federal, state and local (mainly county) components. The federal government provided funds to the states (under a formula based on state level unemployment rates and numbers of eligible persons) and defined the basic outlines

of the program, including eligibility criteria, the basics of program services and operation, and the structure of the performance management system that provided budgetary incentives to local "Service Delivery Areas" (SDAs) that met or achieved certain targets. The states filled in the details of the performance management system and divided up the funds among the SDAs (using the same formula). The local SDAs operated the program on a daily basis, including determining participant eligibility, contracting with local service providers (which included, among others, community organizations, public community colleges and some for-profit providers) and determining, via caseworker consultation with each participant, the assignment to particular services. The performance management system provided an incentive for SDAs to "cream skim" the more employable among their eligible populations into their programs. See Heckman et al. (2002) and Courty and Marschke (2004) for more on the JTPA performance system.

In thinking about participation in JTPA, differences between the U.S. and typical European social safety nets matter. Workers in the U.S. receive Unemployment Insurance (UI) for up to 6 months if they lose their job and have sufficient recent employment. Participation in JTPA does not lengthen UI eligibility. Single parents (and in some cases couples with children and both parents unemployed) can receive cash transfers. Other able bodied adults generally receive only food stamps plus, in some states, cursory cash transfers in the form of general assistance.

The wealth of other programs available providing similar services to those offered by JTPA matters for the interpretation of the participation and non-participation states defined more formally subsequently. Many other government entities at the federal, state and local levels offer job search assistance or classroom training, as do many non-profit organizations. Individuals can also take courses at public 2-year colleges with relatively low tuition (and may also receive government grants or loans to help them do so). As a result of this institutional environment, many control group members receive training services that look like those that treatment group members receive from JTPA. Some of the eligible non-participant comparison group members do as well, with a handful also participating in JTPA itself during the follow-up period.

3.2 The NJS evaluation design

As described in Doolittle and Traeger (1990), the evaluation took place at a non-random sample of only sixteen of the more than 600 JTPA SDAs. Eligible applicants at each site during all or part of the period November 1987–September 1989 were randomly assigned to experimental treatment and control groups. The treatment group remained eligible for JTPA while the control group was embargoed from participation for 18 months. Bloom et al. (1997) summarize the design and findings.

Potential participants received service recommendations prior to random assignment. These recommendations form the basis for the three experimental

"treatment streams" that we analyze in our empirical work. Individuals recommended to receive CT-OS, perhaps in combination with additional services such as JSA but not including OJT, constitute the CT-OS treatment stream. Similarly, individuals recommended to receive OJT, possibly in combination with additional services other than CT-OS, constitute the OJT treatment stream. The residual "Other" treatment stream includes individuals not recommended for either CT-OS or OJT (along with a small number recommended for both). Placing the recommendations prior to randomization allows the estimation of experimental impacts for sub-groups likely to receive particular services. Our analysis focuses on treatment streams because the design just described implies that we have an experimental benchmark for the treatment streams but not for individual treatments.

The NJS also includes a non-experimental component designed to allow the testing of non-experimental evaluation estimators as in LaLonde (1986) and Heckman and Hotz (1989). To support this aspect of the study, data on Eligible Non-Participants (ENPs) were collected at 4 of the 16 experimental sites — Corpus Christi, TX; Fort Wayne, IN; Jersey City, NJ; and Providence, RI. We focus (almost) exclusively on these four SDAs in our empirical analyses.

3.3 Data from the NJS

The data we use come from surveys administered to the ENPs and to controls at the same four sites. These surveys include a long baseline survey, administered shortly after random assignment (RA) for the controls and shortly after measured eligibility (EL) for the ENPs, and follow-up surveys (one or two for the controls and one for the ENPs). Heckman and Smith (1999, 2004) describe the data sets and the construction of these variables in greater detail.

The data on earnings and employment outcomes come from the follow-up surveys. In particular, for the bias estimates we use the same quarterly self-reported earnings variables used in Heckman et al. (1997) and Heckman et al. (1998a). The variables measure earnings (or employment, defined as non-zero earnings) in quarters relative to the month of RA for the controls and of EL for the ENPs. In our work, we aggregate the six quarters after RA/EL into a single dependent variable. Appendix B of Heckman et al. (1998a) describes the construction of these variables, and the resulting analysis sample, in detail. We focus on these variables in order to make our results comparable to those in earlier studies. Our experimental impact estimates use the same earnings variables as in the official impact reports by Bloom et al. (1993) and Orr et al. (1994). These variables differ in a variety of ways; see those reports as well as Heckman and Smith (2000) for more details. Both the JTPA program and the NJS divide the population into four groups based on age and sex: adult males and females aged 22 and older and male and female youth aged 16–21. We focus solely on the two adult groups in this study as they provide the largest samples.

Table 1 displays the sample sizes for our analyses, divided into ENPs, all controls and controls in each of the three treatment streams. Two main points

Table 1 Sample sizes used in estimation

	ENP[a]	All CTRLs	CT-OS	OJT	Other
Adult males					
Propensity score sample[b]	818	734	75	374	285
Observations with non-missing earnings[c]	391	499	57	277	165
After imposing min–max common support		391	49	207	102
Observations with non-missing employment[c]	412	502	57	279	166
After imposing min–max common support		394	49	209	103
Adult females					
Propensity score sample[b]	1,569	869	265	341	263
Observations with non-missing earnings[c]	870	660	207	271	182
After imposing min–max common support		640	200	255	178
Observations with non-missing employment[c]	896	665	208	274	183
After imposing min–max common support		645	201	258	179

[a] No ENP observations are lost due to imposing a common support restriction

[b] The propensity score sample consists of all individuals aged 22 to 54 who completed the long baseline survey and have valid values of the age and sex variables. This is the same sample employed in Heckman and Smith (1999). The sub-samples of the propensity score samples with non-missing values of employment and earnings in the six quarters before and after RA/EL are used in estimating the biases

[c] The sample size for cell matching on the labor force status transitions is slightly smaller than shown here as we cannot use observations with (fractional) imputed values for the transitions. The sample sizes for some of the other estimators are slightly smaller than shown here because the cross-validation sometimes chooses a particular kernel that implicitly imposes a stronger common support restriction

emerge from Table 1: first, our sample sizes, though respectable in comparison to the widely used data from the National Supported Work Demonstration, remain small given that we apply semi-parametric estimation methods. Second, treatment stream assignment does not happen at random. In our data, streams related to services that imply immediate job placement, namely the OJT and "other" streams, have relatively many men, while CT-OS has relatively many women. See Kemple et al. (1993) for a detailed descriptive analysis of assignment to treatment stream in the NJS as a whole.

Table 2 indicates the fraction of the experimental *treatment* group receiving each JTPA service type at the four sites; note that individuals may receive multiple services. Quite similar patterns appear for the full NJS treatment group. These data indicate the extent to which the treatment streams correspond to particular services and aid in the interpretation of the experimental impact estimates presented subsequently. The table highlights two main patterns. First, treatment stream assignment predicts receipt of the corresponding service. For example, among adult women in the CT-OS treatment stream, 58.5% receive some CT-OS, compared to 2.3% in the OJT stream and 10.6% in the "other" stream. Second, as analyzed in detail in Heckman et al. (1998c), many treatment group members, especially those in the OJT and "other" treatment streams, never enroll in JTPA and receive a service. Some treatment group members

Table 2 Treatment streams and service receipt – Percentage of treatment group members receiving each service type: four ENP sites

Actual services received	Experimental treatment stream			
	Overall	CT-OS	OJT	Other
Adult males				
None	44.33	33.94	48.44	41.95
CT-OS	8.27	57.80	0.94	2.97
OJT	14.90	0.92	24.17	6.64
JSA	27.52	21.10	33.85	20.90
ABE	2.86	6.88	0.10	5.37
Others	12.25	3.67	0.42	30.93
Adult Females				
None	46.02	30.16	52.94	52.91
CT-OS	22.02	58.52	2.27	10.55
OJT	9.05	0.16	19.79	5.05
JSA	26.94	26.89	32.09	21.10
ABE	4.92	10.49	0.53	4.74
Others	5.91	3.11	0.53	14.68

Notes:
1 The experimental treatment streams are defined as follows, based on service recommendations prior to random assignment:

The CT-OS stream includes persons recommended to receive CT-OS,
possibly along with services other than OJT, prior to random assignment.
The OJT stream includes persons recommended to receive OJT,
possibly along with services other than CT-OS, prior to random assignment.
The Other services stream includes everyone else.

2 The proportions for actual services received do not have to sum to one because individuals can receive multiple services. These services are:

None is for individuals who do not receive any treatment (drop-outs).
CT-OS is classroom training in occupational skills.
OJT is on-the-job training.
JSA is job search assistance.
ABE is adult basic education.
Other is a mix of other services

received limited services but did not enroll (for reasons related to the gaming of the JTPA performance management system); Table 2 includes only enrollees.

Only a handful of control group members overcame the experimental protocol and received JTPA services in the 18 months after random assignment. At the same time, many control group members, particularly in the CT-OS treatment stream, did receive substitute services from other sources. On average, these services started later than those received by treatment group members and included fewer hours. Exhibits 5.1 and 5.2 of Orr et al. (1994) document the extent of control group substitution for the full NJS; the fraction receiving services in the treatment group exceeds that in the control group by 15–30 percentage points depending on the demographic group and treatment stream. These exhibits combine administrative data on service receipt for the treatment group with self-reports for the control group. Smith and Whalley (2006)

compare the two data sources. See also Heckman et al. (2000), who re-analyze the CT-OS treatment stream data to produce estimates of the impact of training versus no training.

Table 3 presents descriptive statistics on the variables used in the propensity score estimation. Table A1 provides variable definitions. One important variable, namely the labor force status transitions, requires some explanation. A labor force status consists of one of "employed", "unemployed" (not employed but looking for work) and "out of the labor force" (OLF — not employed and not looking for work). Each transition consists of a pair of statuses. The second is always the status in the month of RA/EL. The first is the most recent prior status in the 6 months before RA/EL. Thus, for example, the transition "emp → unm" indicates someone who ended a spell of employment in the 6 months prior to RA/EL to start a spell of unemployment that continued through the month of RA/EL. Transitions with the same status on both sides, such as "unm → unm" correspond to individuals who maintain the same status for all 7 months up to and including the month of RA/EL.

The descriptive statistics reveal a number of interesting patterns. Dropouts (those in the first two education categories) differentially sort into OJT among men but into CT-OS and other among women. Overall, the controls have more schooling than the ENPs. Among adult women, long-term welfare recipients (those in the last welfare transition category) differentially sort into CT-OS, while those not recently on welfare (and in the first transition category) differentially sort into OJT and other. In both groups, individuals unemployed at RA/EL, especially those recently employed or persistently unemployed, differentially sort into the control group; within this group, among men the recent job losers differentially sort into the OJT stream.

4 Econometric methods

4.1 Notation and parameters of interest

This section defines our notation and describes the parameters of interest for the empirical portion of our study. We proceed in the context of the potential outcomes framework variously attributed to Neyman (1923), Fisher (1935), Roy (1951), Quandt (1972) and Rubin (1974). Imbens (2000) and Lechner (2001) extend this framework to multi-treatment programs. Within this framework, we can think about outcomes realized in counterfactual states of the world in which individuals experience treatments they did not receive in real life.

We denote individuals by "i" and treatments by "j" with Y_{ij} signifying the potential outcome for individual "i" in treatment "j". In many multi-treatment program contexts (including ours), it makes sense to single out one treatment as the "no treatment" baseline, which we assign the value $j = 0$. Let $D_{ij} \in \{0, 1\}$ be treatment indicators for each of the $j = 0, \ldots, J$ treatments, where $D_{ij} = 1$ if individual "i" receives treatment "j" and $D_{ij} = 0$ otherwise, where of necessity $\sum_{j=0}^{J} D_{ij} = 1$ for all "i". The observed outcome then becomes $Y_i = \sum_{j=0}^{J} D_{ij} Y_{ij}$.

Table 3 Descriptive statistics

	Adult males				Adult females			
	ENP	CT-OS	OJT	Other	ENP	CT-OS	OJT	Other
Mean age	34.26	29.63	31.99	32.06	33.65	30.26	31.88	32.84
Education								
< 10years	31.76	13.70	25.14	15.44	33.76	23.85	19.70	23.95
10–11 years	17.59	21.92	21.55	27.21	18.91	18.85	19.70	21.67
12 years	29.66	38.36	33.70	36.03	33.56	40.38	47.27	34.60
13–15 years	13.39	21.92	15.47	17.65	10.87	15.38	11.52	15.97
>15 years	7.61	4.11	4.14	3.68	2.90	1.54	1.82	3.80
Race								
White	38.38	17.33	64.71	39.30	37.99	20.75	58.36	36.50
Black	11.74	36.00	19.79	41.75	19.28	35.09	23.17	39.16
Hispanic	44.19	38.67	14.17	14.04	38.12	41.89	15.25	22.81
Other	5.69	8.00	1.34	4.91	4.61	2.26	3.23	1.52
Marital status								
Single	26.17	65.28	43.02	56.55	33.50	56.25	30.89	43.95
Living with spouse	68.60	20.83	36.47	28.84	51.98	19.58	28.03	22.87
Div./ wid./ separated	5.23	13.89	20.51	14.61	14.52	24.17	41.08	33.18
Family income last year								
0–$3,000	16.59	31.71	28.81	42.01	46.48	60.10	38.06	50.52
$3,000–$9,000	17.26	34.15	28.81	23.08	20.02	20.69	34.41	26.80
$9,000–$15,000	21.68	14.63	20.16	19.53	14.45	10.84	14.17	9.28
>$15,000	44.47	19.51	22.22	15.38	19.04	8.37	13.36	13.40
Welfare transition patterns								
No welf. → no welf.	60.17	72.00	75.67	76.14	44.50	33.21	47.80	40.30
No welf. → welfare	1.45	12.00	8.56	7.02	1.64	7.92	13.20	11.41
Welfare → no welf.	1.09	4.00	1.07	2.46	1.71	1.89	3.52	1.90
Welfare → welfare	13.80	9.33	13.64	11.23	36.98	56.60	34.60	44.87
Indicator for missing welfare info.	23.49	2.67	1.07	3.16	15.17	0.38	0.88	1.52
Labor force transition patterns								
emp → emp	70.22	14.29	20.99	18.83	36.58	15.73	19.46	13.62
unm → emp	6.99	11.11	13.27	8.79	4.09	2.02	8.72	9.36
olf → emp	2.50	4.76	4.94	5.02	4.50	1.61	5.37	4.68
emp → unm	5.16	28.57	28.40	27.20	3.76	12.50	24.16	18.72
unm → unm	4.99	23.81	17.90	13.81	4.09	16.94	12.42	14.04
olf → unm	1.33	7.94	3.09	7.95	3.85	8.47	9.40	10.64
emp → olf	1.50	3.17	6.79	5.02	5.65	7.66	7.72	4.26
unm → olf	0.33	1.59	2.16	1.67	2.37	5.24	2.68	4.26
olf → olf	6.99	4.76	2.47	11.72	35.11	29.84	10.07	20.43
Sum of earnings								
6 pre-RA/EL quarters	16838.7	9607.7	10401.6	9857.0	6096.8	4276.1	6795.8	5308.4
6 post-RA/EL quarters	18902.1	9975.9	13196.9	12037.4	7112.7	5750.3	9131.7	7213.2
Employed								
In quarter 6 before RA/EL	0.759	0.564	0.680	0.680	0.454	0.416	0.534	0.488
In quarter 6 after RA/EL	0.736	0.667	0.703	0.675	0.498	0.486	0.672	0.552

Note: The descriptive statistics apply to the sample used to estimate the propensity scores

In our data $j = 1$ denotes the CT-OS treatment stream, $j = 2$ denotes the OJT treatment stream and $j = 3$ denotes the other treatment stream. To reduce notational burden we omit the "j" subscript when it is not needed. Within treatment stream "j", individuals randomly assigned to the experimental treatment state experience Y_{ij} and those randomly assigned to the control group (along with the ENPs) experience Y_{i0}. These states embody both failure to enroll in JTPA in the first case and possible service receipt from other programs (by both the controls and the ENPs) in the second case.

The most common parameter of interest in the literature consists of the average impact of treatment "j" on the treated, given by

$$\text{ATET}_j = E(Y_j|D_j = 1) - E(Y_0|D_j = 1).$$

This parameter indicates the mean effect of receiving treatment "j" relative to receiving no treatment for those individuals who receive treatment "j". The average treatment effect on the treated for the multi-treatment program as a whole consists of a weighted (by the fraction in each treatment) average of the ATET_j.

We have the rich data on conditioning variables required to justify the matching methods we use only for the experimental controls and the ENPs. As a result, rather than estimating average treatment effects, we follow Heckman et al. (1997, 1998a) in estimating the bias associated with applying matching based on covariates X to these data to estimate $\text{ATET}_j, j \in \{1, 2, 3\}$. For treatment stream "$j$" this bias equals

$$\text{BIAS}_j = \int \left[E(Y_{0i}|X_i, D_{ij} = 1) - E(Y_{0i}|X_i, D_{i0} = 1)\right] \mathrm{d}f(X|D_{ij} = 1),$$

where the first term inside the square brackets corresponds to the experimental control group for treatment stream "j" and the second term corresponds to the ENPs. Integrating with respect to the distribution of observables for the control group reflects our interest in the bias associated with estimating the ATET. If $\text{BIAS}_j = 0$ then matching using conditioning variables X solves the selection problem in this context for treatment stream "j". In essence, we view each treatment stream as a separate program and estimate the bias associated with using matching to estimate the ATET for that treatment stream using the ENPs as a comparison group.

The literature defines a variety of other parameters of interest. The unconditional average treatment effect, defined as $\text{ATE}_j = E(Y_j) - E(Y_0)$, provides useful information when considering assigning all of some population to a particular treatment. In a multi-treatment program context, Imbens (2000) and Lechner (2001) define a variety of other parameters, such as the mean impact of receiving treatment "j" relative to treatment "k" for those who receive treatment "j" and the mean impact of treatment "j" on those who receive either treatment "j" or treatment "k". Due to the nature of our data we do not examine

these additional parameters, nor do we use the more complicated apparatus of multi-treatment matching developed by Imbens (2000) and Lechner (2001).

Moreover, all of the parameters defined in this section represent partial equilibrium parameters, in the sense that they treat the potential outcomes as fixed when changing treatment assignment. The statistics literature calls this the Stable Unit Treatment Value Assumption (SUTVA). Heckman et al. (1998b), Lise et al. (2005) and Plesca (2006) discuss program evaluation in a general equilibrium context.

4.2 Identification

Our empirical analysis follows the literature that treats JTPA as a single treatment by using the experimental data as a benchmark against which to judge the performance of semi-parametric matching estimators. We use matching for four reasons. First, it performs reasonably well in the existing literature at evaluating the aggregated JTPA treatment. Second, we have very rich data on factors related to participation and outcomes, including monthly information on labor force status in the period prior to the participation decision. The existing literature, in particular Card and Sullivan (1988), Heckman and Smith (1999) and Dolton et al. (2006) emphasizes both the importance of conditioning on past labor market outcomes and doing so in flexible ways. Third, relative to least squares regression, matching only compares the comparable when constructing the estimated, expected counterfactual, allows for more flexible conditioning on the observables and allows an easier examination of the support condition. Fourth, while this does not make matching any more plausible, we lack the exclusion restrictions required to use IV or the bivariate normal selection model of Heckman (1979). Furthermore, Heckman and Smith (1999) find, for reasons discussed below, that longitudinal estimators fare poorly in this context.

Matching estimators of all sorts rely on the assumption of selection on observables; that is, they assume independence between treatment status and untreated outcomes conditional on some set of observable characteristics. In the matching literature, this gets formalized as the conditional independence assumption (CIA), $Y_0 \perp D|X$, where "$\perp$" denotes independence. The statistics literature calls this assumption unconfoundedness. As noted in Heckman et al. (1997, 1998a) our problem actually requires only mean independence, rather than full independence. We invoke the CIA separately for each of the three treatment streams.

Rosenbaum and Rubin (1983) show that if you can match on some set of conditioning variables X, then you can also match on the probability of participation given X, or the propensity score, given by $P(X) = \Pr(D = 1|X)$. Their finding allows the restatement of the CIA in terms of $P(X)$. Matching (or weighting) on estimated propensity scores from a flexible parametric propensity score model reduces the non-parametric dimensionality of the problem from the number of conditioning variables to one, thus substantially increasing the rate of convergence. Use of a flexible parametric propensity score model

seems to perform as well in practice as either reducing the dimensionality of X via alternative means such as the Mahalanobis metric or estimating propensity scores semi-parametrically. See Zhao (2004) for further discussion of alternative dimension reduction schemes and Kordas and Lehrer (2004) for a discussion of semi-parametric propensity scores.

In order for the CIA to have empirical content, the data must include untreated observations for each value of X observed for a treated observation. In formal terms, in order to estimate the mean impact of treatment on the treated, we require the following common support condition: $P(X) < 1$ for all X. This condition can hold in the population, or in both the population and the sample, though the literature often neglects this distinction. We assume it holds in the population and then impose it in the sample. As discussed in e.g. Smith and Todd (2005a), a number of methods exist to impose this condition. We adopt the simple min-max rule employed in Dehejia and Wahba (1999, 2002); under this rule, observations below the maximum of the two minimums of the estimated propensity scores in the treated and untreated samples, and above the minimum of the maximums, lie outside the empirical common support and get omitted from the analysis. We adopt this rule rather than the more elegant trimming rule employed in Heckman et al. (1997, 1998a) for simplicity given that our sensitivity analysis reveals no substantive effect of this choice (or, indeed, of simply ignoring the issue) on the results. Given our focus on pairwise comparisons between treatment types and no treatment, we apply the support condition separately for each pairwise comparison.

4.3 Estimation

We estimate our propensity scores using a standard logit model. The only twist concerns adjustment for the choice-based sampling that generated our data. Our data strongly over-represent participants relative to their fraction in the population of JTPA eligibles. We follow Heckman and Smith (1999) in dealing with this issue by reweighing the logit back to population proportions under the assumption that controls represent three percent of the eligible population; see their footnote 19 for more on this. We further assume that each treatment stream represents one percent of the eligible population.

Smith and Todd (2005b) show that the literature offers a variety of alternative balancing tests. These tests aid the researcher in selecting an appropriately flexible parametric propensity score model for a given set of conditioning variables X by examining the extent to which a given specification satisfies the property that $E(D|X, P(X)) = E(D|P(X))$. In words, conditional on $P(X)$, the X should have the same distribution in the treated and comparison groups. In this sense, matching mimics a randomized experiment by balancing the distribution of covariates in the treatment group and the matched (or reweighted) comparison group. Balancing tests do not provide any information about the validity of the CIA. For simplicity and comparability with most of the existing literature, we focus here on the "standardized differences" described in

Rosenbaum and Rubin (1985). For each variable in X, the difference equals the mean in the treatment group minus the mean in the matched (or reweighted) comparison group divided by the square root of the sum of the variances in the treated and unmatched comparison groups. Rosenbaum and Rubin (1985) suggest concern in regard to values greater than 20.

As one of the results from the existing literature that we want to revisit in the disaggregated context concerns a general lack of sensitivity to the particular matching estimator selected, we report estimates from a number of different matching estimators here, along with OLS and two cell matching estimators. All matching estimators have the general form

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

where n_1 denotes the number of $D = 1$ observations. They differ only in the details of the construction of the weight function $w(i,j)$. As described in e.g. Angrist and Krueger (1999), OLS also implicitly embodies a set of weights that, depending on the distributions of X among the participants and non-participants, can differ substantially from those implied by most matching estimators.

We can also think about matching as using the predicted values from a non-parametric regression of Y_0 on $P(X)$ estimated using the comparison group sample as the estimated, expected counterfactual outcomes for the treated units. This way of thinking about matching makes it clear both that matching differs less from standard methods than it might first appear and that all our knowledge about various non-parametric regression methods, such as that in Pagan and Ullah (1999), applies in this context as well. Each matching method we consider, with the exception of the longitudinal ones, corresponds to using a different estimator for the non-parametric regression of Y_0 on $P(X)$.

We consider two simple cell matching estimators. The first matches observations solely on the value of their labor force status transition variable. The second estimator stratifies based on deciles of the estimated propensity score, where the deciles correspond to the pooled sample. The applied statistics literature often uses this approach, though that literature often follows Rosenbaum and Rubin (1984) in using only five propensity score strata. As we are cautious economists rather than bold statisticians, we use 10 in our analysis.

In nearest neighbor matching $w(i,j) = 1$ for the comparison observation that has the propensity score closest to that of treated observation "i" and zero otherwise. We implement nearest neighbor matching with replacement, so that a given comparison observation can get matched to more than one treated observation, because our data suffer from a lack of comparison group observations similar to the treated observations.

Kernel matching assigns positive weight to comparison observations with propensity scores similar to that of each treated observation, where the weights decrease with the propensity score distance. Formally,

$$w(i,j) = \frac{G\left(\frac{P_i(X)-P_j(X)}{a_n}\right)}{\sum_{k\in\{D_j=0\}} G\left(\frac{P_i(X)-P_k(X)}{a_n}\right)},$$

where G denotes a kernel function and a_n denotes an appropriately chosen bandwidth. We consider three commonly used kernels: the Gaussian (the standard normal density function), the Epanechnikov and the tricube. Local linear matching uses the predicted values from a local linear regression (a regression weighted by the kernel weights just defined) as the estimated expected counterfactual. Fan and Gijbels (1996) discuss the relative merits of kernel regression versus local linear regression; for our purposes, the fact that local linear regression has better properties near boundary values suggests applying it here, given our many observations with propensity scores near zero.

Though not required for consistency, ex post regression adjustment following matching — essentially running a regression using the weights from the matching — can reduce bias in finite samples and also reduce the variance of the resulting estimate. The formal literature calls this bias-corrected matching. See Ho et al. (2007) for informal discussion, references and applications. Note that this procedure differs from the "regression-adjusted" matching in Heckman et al. (1997, 1998a) because here the matching step comes first.

Finally, in addition to cross-sectional matching estimators, we consider two variants of the difference-in-differences matching developed in Heckman et al. (1997, 1998a). This method differs from standard differences-in-differences because it uses matching rather than linear regression to condition on X. We simply replace the post-RA/EL outcome measure with the pre–post difference to implement the estimator.

Each class of matching estimators (other than cell matching) implies a bandwidth choice. Choosing a wide bandwidth (or many neighbors in the nearest neighbor matching) reduces the variance of the estimates because more observations, and thus more information, go into the predicted expected counterfactual for each observation. At the same time, a wider bandwidth means more bias, as observations less like the treated observation under consideration get used in constructing the counterfactual. In our analysis, we allow the data to resolve the matter by relying on leave-one-out cross validation as described in e.g. Racine and Li (2005) and implemented in Black and Smith (2004) to choose bandwidths that minimize the estimated mean squared error of the estimates. Fitzenberger and Speckesser (2005) and Galdo et al. (2006) consider alternative bandwidth selection schemes.

In the kernel matching, we also rely on the cross-validation to choose among the Gaussian, Epanechnikov and tricube kernel functions. As the second and third of these do not imply positive weights on the whole real line, they may implicitly strengthen the support condition we impose. As we use the same ENP comparison group when analyzing each treatment stream, we need only one bandwidth for each estimator for each demographic group. Table A2 documents the bandwidth choice exercise.

Heckman and Todd (1995) consider the application of matching methods in choice-based samples (such as ours). Building on the robustness of logit model coefficient estimates (other than the intercept) to choice-based sampling, they show that matching works in choice-based samples when applied using the odds ratio or the log odds ratio from an unweighed logit participation model. Theory provides no guidance on whether to use the odds ratio or the log odds ratio; as we have many estimated scores near zero, we use the odds ratio to better distinguish these values. In any event, a sensitivity analysis revealed little effect of this decision on the estimates.

5 Empirical analysis of the NJS

5.1 Experimental estimates

We begin our empirical analysis by looking for the possibility of cancellation when combining impacts from the three treatment streams in the NJS data. Given that the different services (and thus the different treatment streams) involve quite different inputs in terms of time and other resources — see e.g. the cost estimates in Exhibits 6.4 and 6.5 of Orr et al. (1994) and Heinrich et al. (1999) — and given the use of different providers for the various services within JTPA, we have good reasons to expect differences in mean impacts by treatment stream.

Table 4 reports experimental impact estimates over 18 and 30 months after random assignment, respectively, for both adult males and adult females. The impacts at 18 months in Table 4 are based solely on self-reported earnings from the first follow-up survey, with outliers recoded by hand by Abt Associates — the same outcome variable as in Bloom et al. (1993). The impacts at 30 months in Table 4 rely on the earnings variables from Orr et al. (1994), which combine self-reported data from both follow-up surveys with administrative data from state UI records for non-respondents in a rather unattractive way (see their Appendix A for the sordid details). We define employment as non-zero earnings in the sixth or tenth quarters after random assignment. All estimates consist of simple mean differences. Heckman and Smith (2000) analyze the sensitivity of the NJS experimental impact estimates.

Table 4 reveals four important patterns. First, the impact estimates have non-trivial standard errors; conditioning on observables would not change this very much. Not surprisingly, we typically find smaller standard errors for all 16 sites than for the four ENP sites. Second, the point estimates vary a lot by treatment stream. Although not close to statistical significance at 18 months, at 30 months the employment estimates for adult females do statistically differ by treatment stream. Moreover, for both the four and 16 site estimates, three of the four comparisons have p values below 0.20. This suggests the potential for substantively meaningful cancellation when, for example, combining the strong employment impacts in quarter 10 for adult women in the CT-OS stream with the zero estimated impact for those in the OJT stream.

Table 4 Adult males and females - experimental impacts by treatment stream

	Overall	CT-OS	OJT	Other	Test of equality across streams [b]
Experimental impacts at 18 months					
Impacts at four sites with ENPs [a]					
Outcome: sum of earnings over 18 months after random assignment					
Adult males	427.66	36.82	1264.86	−1228.18	Chi2(2) =2.71
	(651.61)	(1513.09)	(841.77)	(1285.94)	p-value = 0.26
Adult females	473.99	335.21	847.85	65.73	Chi2(2) = 0.56
	(424.48)	(642.33)	(652.45)	(937.71)	p-value = 0.75
Outcome: employment in quarter 6 after random assignment					
Adult males	0.011	−0.032	0.040	−0.033	Chi2(2) = 2.06
	(0.025)	(0.072)	(0.032)	(0.047)	p-value = 0.36
Adult females	0.025	0.038	0.016	0.027	Chi2(2) = 0.20
	(0.022)	(0.041)	(0.031)	(0.044)	p-value = 0.91
Impacts at all 16 experimental sites					
Outcome: sum of earnings over 18 months after random assignment					
Adult males	572.89	397.79	831.08	297.95	Chi2(2) = 0.41
	(381.04)	(745.08)	(525.26)	(809.16)	p-value = 0.82
Adult females	765.48	700.93	735.40	1047.86	Chi2(2) = 0.30
	(230.54)	(318.94)	(392.40)	(561.26)	p-value = 0.86
Outcome: employment in quarter 6 after random assignment					
Adult males	0.016	0.005	0.021	0.015	Chi2(2) = 0.18
	(0.015)	(0.030)	(0.020)	(0.030)	p-value = 0.92
Adult females	0.030	0.034	0.010	0.059	Chi2(2) = 2.06
	(0.013)	(0.020)	(0.020)	(0.028)	p-value = 0.36
Experimental Impacts at 30 Months					
Impacts at four sites with ENPs [a]					
Outcome: sum of earnings over 30 months after random assignment					
Adult males	942.01	−1272.44	1964.79	−495.47	Chi2(2) = 2.20
	(897.57)	(3008.57)	(1256.27)	(1376.23)	p-value = 0.33
Adult females	1565.85	756.99	883.45	3288.96	Chi2(2) = 3.51
	(627.93)	(1082.29)	(988.27)	(1094.15)	p-value = 0.17
Outcome: employment in quarter 10 after random assignment					
Adult males	−0.030	−0.090	0.006	−0.086	Chi2(2) = 4.25
	(0.023)	(0.080)	(0.030)	(0.037)	p-value = 0.12
Adult females	0.054	0.087	0.000	0.108	Chi2(2) = 5.20
	(0.022)	(0.045)	(0.033)	(0.039)	p-value = 0.07
Impacts at all 16 experimental sites					
Outcome: sum of earnings over 30 months after random assignment					
Adult males	1213.22	1266.72	1675.36	388.44	Chi2(2) = 0.91
	(580.94)	(1245.81)	(829.46)	(1065.74)	p-value = 0.63
Adult females	1248.79	912.66	749.88	2638.24	Chi2(2) = 4.32
	(369.52)	(548.18)	(633.38)	(761.96)	p-value = 0.12
Outcome: employment in quarter 10 after random assignment					
Adult males	0.007	−0.005	0.033	−0.036	Chi2(2) = 4.05
	(0.015)	(0.033)	(0.021)	(0.027)	p-value = 0.13
Adult females	0.037	0.037	0.013	0.078	Chi2(2) = 3.23
	(0.014)	(0.022)	(0.022)	(0.028)	p-value = 0.20

[a] Robust standard errors are in parentheses

[b] The null hypothesis is equal impacts in the three treatment streams

5.2 Determinants of participation by treatment stream

All of the services offered by JTPA aim to improve the labor market prospects of participants. At the same time, the channels through which they operate, and the economics of the participation decision related to each service, differ substantially. For example, CT-OS represents a serious investment in human capital that aims to prepare the participant for a semi-skilled occupation and thereby increase their wage. It has a higher opportunity cost than the other services because participants typically do not work while receiving training and because, unlike many European programs, participants do not receive any stipend (though they remain eligible for other transfers). OJT immediately places the participant in employment. Participants in OJT get a chance at employers who might reject them without the subsidy (which gives employers an incentive to take some risks in hiring, keeping in mind the low dismissal costs in the U.S.) as well as human capital acquired on the job. This service has low opportunity costs but has the feature that not only must the caseworker agree to provide the subsidy but a firm must also agree to hire the subsidized worker. Finally, the Job Search Assistance (JSA) received by many in the "other" services stream aims to reduce the time required to find a job, but does not aim to increase wages via increases in human capital.

Because of these differences in the economics among the services offered by JTPA, we expect the nature of the selection process to differ by treatment stream. These differences may affect the timing and magnitude of the "Ashenfelter (1978) dip." As discussed in Heckman and Smith (1999) and documented for a variety of programs in Heckman et al. (1999), the dip refers to the fall in mean earnings and employment typically observed among participants just prior to participation. These differences may also affect what variables matter, and how strongly they matter, in predicting participation conditional on eligibility. For example, we expect to see job-ready participants, as indicated by past labor force attachment and schooling, receiving OJT, and to see individuals with less human capital and with sources of income from social programs, such as single mothers on AFDC, sort into CT-OS.

We begin by looking at Ashenfelter's dip. Figures 1 and 2 present the time series of mean earnings. Figure 1 shows that (somewhat surprisingly) for adult men all three treatment streams display roughly the same pattern as the full control group in terms of both levels and dip, though with a slightly muted dip for the "other" treatment stream. Figure 2 for adult women shows similar dips across treatment streams, this time slightly magnified for the "other" treatment stream, but different initial levels across groups. Consistent with the earlier discussion, those who enter the CT-OS stream have the lowest earnings levels and those entering the OJT stream have the highest levels, which is suggestive of greater job readiness. The lack of strong differences in the pre-program dip among the treatment streams surprised us. For adult women, we also observe post-random assignment earnings growth relative to the ENPs for all three treatment streams. Heckman and Smith (1999) document that the dip, along with the post-random assignment earnings growth observed for adult female

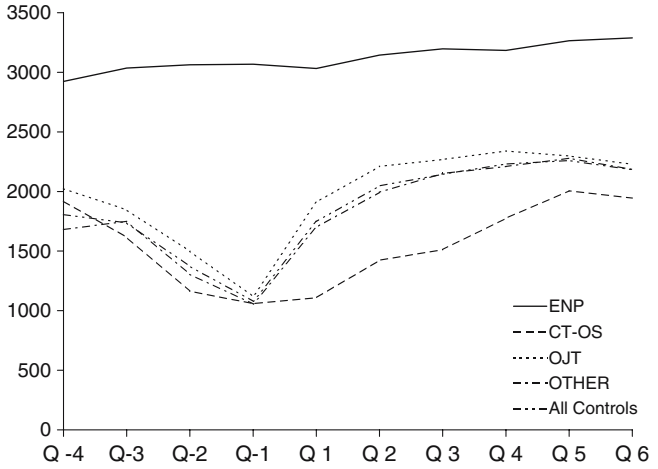

Fig. 1 Adult males – pre-RA/EL and post-RA/EL monthly earnings averaged by quarter

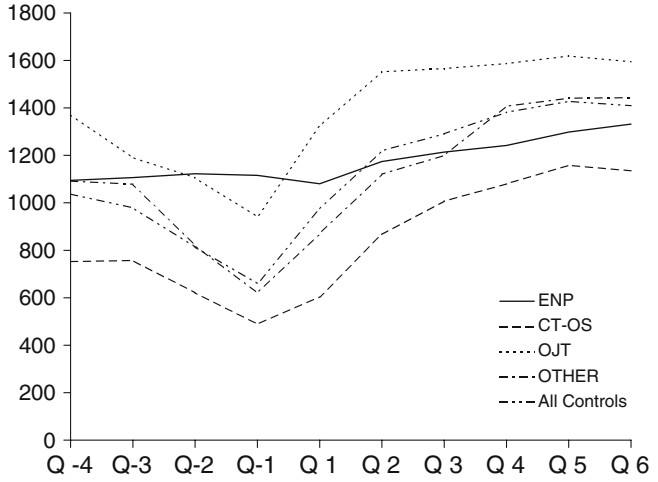

Fig. 2 Adult females – pre-RA/EL and post-RA/EL monthly earnings averaged by quarter

controls, imply both sensitivity to the choice of before and after periods and bias in longitudinal estimators. For this reason, we focus primarily on cross-sectional estimators that condition on lagged labor market outcomes in Sect. 5.3.

Table 5 presents mean derivatives (or finite differences in the case of binary or categorical variables included as a series of indicators) and associated estimated standard errors from logit models of participation in JTPA overall estimated using the full control group and the ENPs along with similar models for each treatment stream estimated using the controls from that stream. The table also presents the *p* values from tests of the joint significance of categorical variables

Table 5 Adult males and females – mean derivatives from logit model of participation

	Overall	CT-OS	OJT	Other
		Males		
Site: Fort Wayne	0.163	−0.020	0.089	0.135
	(0.032)	(0.008)	(0.022)	(0.033)
Site: Jersey City	0.020	0.017	−0.030	0.044
	(0.034)	(0.014)	(0.019)	(0.028)
Site: providence	0.107	0.009	−0.017	0.165
	(0.040)	(0.016)	(0.026)	(0.037)
Test site = 0				
(*p*-values)	0.03	0.71	0.16	0.27
Race: black	0.068	0.012	0.018	0.081
	(0.029)	(0.016)	(0.019)	(0.021)
Race: other[a]	0.000	0.009	−0.023	0.031
	(0.027)	(0.014)	(0.019)	(0.021)
Test race = 0				
(*p*-values)	0.33	0.91	0.81	0.28
Age	−0.019	0.009	−0.016	−0.007
	(0.010)	(0.008)	(0.007)	(0.006)
Age squared	0.0002	−0.0002	0.0002	0.0001
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Test age = 0				
(*p*-values)	0.41	0.65	0.64	0.86
Education <10 years	−0.057	−0.035	−0.010	−0.047
	(0.026)	(0.014)	(0.017)	(0.016)
Education 10–11years	0.015	−0.009	0.017	0.014
	(0.025)	(0.012)	(0.016)	(0.014)
Test education = 0				
(*p*-values)	0.31	0.59	0.87	0.50
Married at RA/EL[b]	−0.062	−0.031	−0.034	−0.020
	(0.022)	(0.013)	(0.016)	(0.014)
Family income				
$3,000–$15,000		0.004		
		(0.019)		
$3,000–$9,000	0.013		0.040	−0.015
	(0.036)		(0.024)	(0.020)
$9,000–$15,000	0.052		0.059	0.018
	(0.038)		(0.027)	(0.021)
> $15,000	−0.052	0.011	0.007	−0.075
	(0.043)	(0.017)	(0.029)	(0.031)
Test family income = 0				
(*p*-values)	0.45	0.98	0.71	0.67
LF transition into unempl.	0.196	0.050	0.123	0.104
	(0.031)	(0.016)	(0.025)	(0.021)
LF transition into OLF	0.070	0.016	0.050	0.045
	(0.037)	(0.019)	(0.029)	(0.024)
Test LF transitions = 0				
(*p*-values)	0.00	0.24	0.03	0.10
(Earnings Q-1)/1,000	−0.023	0.003	−0.022	−0.004
	(0.012)	(0.008)	(0.010)	(0.008)
(Earnings Q-2)/1,000	−0.009	−0.008	0.002	−0.010
	(0.011)	(0.007)	(0.008)	(0.007)
(Earnings Q-3 to Q-6)/1,000	0.006	0.000	0.003	0.005
	(0.002)	(0.000)	(0.001)	(0.001)

Table 5 continued

	Overall	CT-OS	OJT	Other
Test past earnings = 0 (*p*-values)	0.11	0.61	0.56	0.53
Pseudo-*R* square	0.28	0.30	0.28	0.31
		Females		
Site: Fort Wayne	0.050	−0.009	0.018	0.057
	(0.025)	(0.007)	(0.014)	(0.042)
Site: Jersey City	0.020	0.005	−0.004	0.024
	(0.022)	(0.009)	(0.015)	(0.029)
Site: Providence	0.011	0.001	−0.016	0.041
	(0.024)	(0.010)	(0.011)	(0.041)
Test site = 0 (*p*-values)	0.08	0.55	0.13	0.09
Race: black	0.000	0.002	−0.010	0.011
	(0.016)	(0.010)	(0.010)	(0.012)
Race: other[a]	0.013	0.004	−0.006	0.017
	(0.019)	(0.010)	(0.013)	(0.019)
Test race = 0 (*p*-values)	0.75	0.92	0.69	0.44
Welfare trans.	0.091	0.025	0.042	0.031
No welfare → welfare	(0.045)	(0.033)	(0.034)	(0.033)
Welfare trans.	0.010	0.006	0.013	−0.006
Welfare → no welfare	(0.043)	(0.033)	(0.036)	(0.017)
Welfare trans.	−0.005	0.000	−0.003	−0.003
Welfare → welfare	(0.014)	(0.008)	(0.011)	(0.009)
Indicator for missing	−0.051	−0.014	−0.021	−0.013
Welfare information	(0.010)	(0.004)	(0.008)	(0.008)
Test welfare trans. = 0 (*p*-values)	0.01	0.66	0.26	0.44
Age	−0.003	−0.001	−0.002	0.001
	(0.006)	(0.003)	(0.004)	(0.004)
Age squared	0.0000	0.0000	0.0000	0.0000
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
HS dropout	−0.015	−0.005	−0.010	0.001
	(0.013)	(0.007)	(0.010)	(0.008)
Educ. >13 years	0.001	0.000	−0.005	0.006
	(0.017)	(0.009)	(0.014)	(0.010)
Test education = 0 (*p*-values)	0.47	0.73	0.60	0.80
Married at RA/EL[b]	−0.058	−0.015	−0.024	−0.021
	(0.018)	(0.011)	(0.013)	(0.013)
Family income				
$3,000 – $9,000	0.024	0.005	0.015	0.005
	(0.017)	(0.010)	(0.013)	(0.011)
$9,000 – $15,000	0.008	0.007	0.005	−0.002
	(0.025)	(0.014)	(0.018)	(0.017)
>$15,000	0.009	0.005	0.002	0.007
	(0.027)	(0.016)	(0.021)	(0.017)
Test family income = 0 (*p*-values)	0.57	0.94	0.66	0.93
LF transition	0.062	−0.003	0.030	0.032
unm → emp	(0.029)	(0.025)	(0.020)	(0.019)
LF transition	0.039	0.001	0.022	0.016

Table 5 continued

	Overall	CT-OS	OJT	Other
olf → emp	(0.035)	(0.028)	(0.024)	(0.021)
LF transition	0.045	0.011	0.025	0.010
emp → olf	(0.029)	(0.015)	(0.021)	(0.022)
LF transition	0.071	0.018	0.029	0.028
unm → olf	(0.035)	(0.018)	(0.030)	(0.023)
LF transition	0.026	0.009	−0.001	0.068
olf → olf	(0.024)	(0.088)	(0.020)	(0.016)
LF transition	0.103	0.025	0.044	0.037
into unempl.	(0.023)	(0.013)	(0.016)	(0.016)
Test LF transitions = 0 (*p*-values)	0.00	0.26	0.02	0.05
Pseudo-*R* square	0.17	0.15	0.21	0.18

The values in the table are mean derivatives; standard errors are in parentheses

[a] Due to small sample sizes, we combine the "Hispanic" and "other" categories here

[b] RA/EL is the month of random assignment for the experimental controls and the month of measured eligibility for the ENPs

Table 6 Adult males and females — tests of equality of logit coefficients[a]

Adult males		Adult females	
Site	chi2(6) = 26.4 *p*-value = 0.00	Site	chi2(6) = 71.61 *p*-value = 0.00
Race	chi2(4) = 7.03 *p*-value = 0.13	Race	chi2(4) = 16.54 *p*-value = 0.00
Age and age squared	chi2(4) = 1.2 *p*-value = 0.88	Age and age squared	chi2(4) = 9.73 *p*-value = 0.05
Education	chi2(4) = 4.44 *p*-value = 0.35	Education	chi2(4) = 5.91 *p*-value = 0.21
Married at RA/EL[b]	chi2(2) = 0.71 *p*-value = 0.70	Married at RA/EL	chi2(2) = 0.55 *p*-value = 0.76
Family income	chi2(6) = 5.26 *p*-value = 0.51	Family income	chi2(6) = 4.12 *p*-value = 0.66
LF transitions	chi2(4) = 0.12 *p*-value = 1.00	LF transitions	chi2(12) = 18.00 *p*-value = 0.12
Past earnings (last two quarters)	chi2(4) = 1.66 *p*-value = 0.80	Welfare transitions	chi2(8) = 3.64 *p*-value = 0.89

[a] The null hypothesis is equal coefficients across the three treatment streams

[b] RA/EL is the month of random assignment for the experimental controls and the month of measured eligibility for the ENPs

included as a series of indicators. Table 6 presents the chi-squared statistics and related *p* values from tests of the null of equal coefficients across treatment streams for particular variables or categories of variables.

For the matching estimators applied below, we want to include all the variables that affect both participation and outcomes. The specifications presented here differ somewhat from those in Heckman et al. (1997, 1998a) and Heckman and Smith (1999), upon whose analyses we build. Those papers consider

economic theory, institutional knowledge, predictive power and statistical significance as variable selection criteria. Our choices emphasize the knowledge gained from those earlier papers combined with a desire for greater parsimony given the relatively smaller sample sizes available once we split the sample into treatment streams. We considered several less parsimonious specifications and found that they yielded the same general conclusions. Heckman and Navarro (2004) discuss the variable selection issue in greater depth.

For adult men, our final specification includes site and race indicators, age and age squared, education categories, marital status, categories of family income in the year prior to RA/EL, labor force status transitions (collapsed into coarser categories) and own quarterly earnings in quarters prior to RA/EL. The specification for adult women differs in that it includes welfare status transitions but omits the quarterly earnings variables (which matter less for this group) and, because of the larger sample, it does not collapse the labor force status transition categories. We do not worry about the potential endogeneity of the labor force histories for reasons outlined in Frölich (2006). To produce consistent estimates of the treatment effects, we only need to balance the unobservable conditional on X and D not to make the bias zero conditional on X and D; non-parametric regression accomplishes this because it compares, in the limit, only observations with the same X (or the same propensity score).

Balancing test results for all of the cross-sectional matching estimators appear in Table A3. Figures 3 and 4 show the distributions of propensity scores. Consistent with the non-trivial numbers of observations lost when imposing the common support condition in Table 1, we find important support problems for larger values of the scores.

For the full control group, our findings mimic those presented in Heckman and Smith (1999, Table 6) and Heckman et al. (1998a, Table III)). In particular, they replicate the importance of the labor force transition variables for both groups, as well as the welfare transition variable for adult women and pre-RA/EL earnings for adult men. For the individual treatment streams, we find both similarities with the overall results and differences, in addition to a general reduction in precision due to the reduced sample sizes. In particular, we find evidence that the coefficients on the site variables, the race variables and, for women, the labor force status transition variables differ among the three streams. As the sites differ strongly in their relative emphasis on the different treatment types, the first finding comes as no surprise. For both groups, blacks and other non-whites have higher probabilities of assignment to the "other" stream relative to whites and lower probabilities of assignment to the OJT stream. This finding suggests that, conditional on the other covariates, caseworkers, employers or the participants themselves think that non-whites make better candidates for JSA, the most common service in the other stream, and worse candidates for OJT. This could reflect real or perceived discrimination on the part of caseworkers or firms providing OJT positions or it could mean that non-whites more often receive non-JSA services within the other stream.

For adult women, the labor force transitions have much smaller mean derivatives in the CT-OS stream than in the other two streams (the same pattern holds

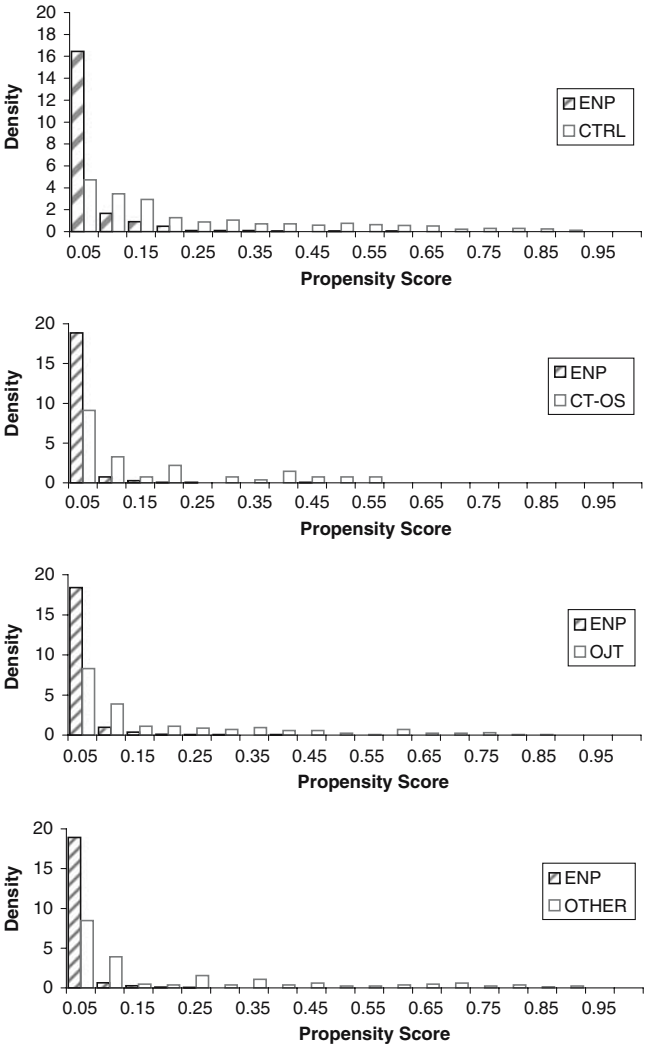

Fig. 3 Adult males — distribution of propensity score

for adult men but does not reach the usual levels of statistical significance). Also, women out of the labor force in the 7 months up to and including RA/EL have much higher mean probabilities of participation, relative to women employed during those months, in the other treatment stream than in the CT-OS and OJT treatment streams. The labor force transition findings suggest that these variables contain information about the individual's readiness for, and eagerness to obtain, employment; thus, they matter for the OJT and other streams, whose members typically receive OJT or JSA, both of which aim at immediate placement. Put differently, for this group, distinguishing among sets of individuals all of whom have zero earnings in the month of RA/EL (which means six of the nine transition categories) matters for adult women in two of the streams,

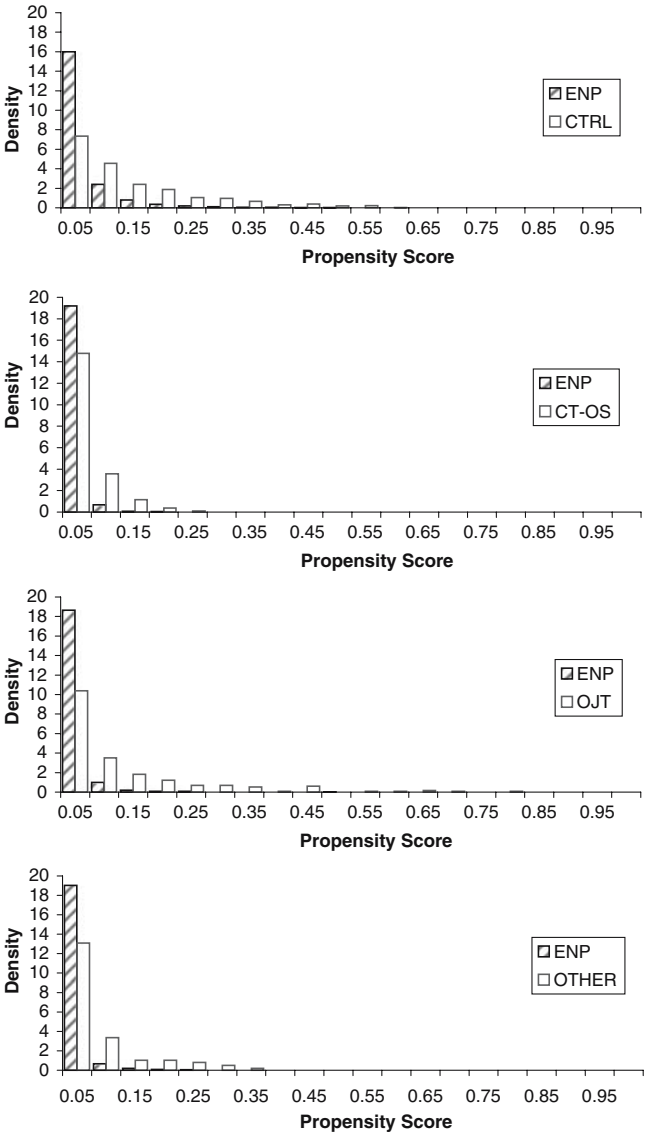

Fig. 4 Adult females — distribution of propensity score

and reinforces the value of collecting information on labor force status at a fine level of temporal detail.

Overall, the differential effects of site, race and labor force status transitions across the three treatment streams represent important and interesting findings. These results enrich our view of how JTPA operated, suggest hypotheses to test in future evaluations of other multi-treatment programs and illustrate the potential knowledge gain associated with separately examining individual treatments within multi-treatment programs.

5.3 Selection bias and the performance of matching estimators

Table 7 presents bias estimates, along with bootstrap standard errors, for the matching estimators described in Sect. 4.3. We also present the estimated Root Mean Squared Error (RMSE) associated with each estimator, defined as:

$$\text{RMSE} = \sqrt{\text{var}(\widehat{BIAS}) + \widehat{BIAS}^2}.$$

Given the large number of estimators we examine and the computational burden associated with bootstrapping, we limited ourselves to only 50 bootstrap replications, a number likely well below that implied by the analysis in Andrews and Buchinsky (2000); the reader should thus keep in mind that the variances themselves represent noisy estimates. For simplicity, we present bootstrap standard errors for all of the estimators other than OLS and the cell matching estimators, despite the problems with doing so in the case of the nearest neighbor estimators laid out in Abadie and Imbens (2006). Fortunately, their Monte Carlo analysis suggests that use of the bootstrap does not lead to severely misleading inferences.

The outcome variables consist of the sum of earnings in the 18 months after random assignment and employment in quarter six after random assignment. Recall that we estimate biases, not average treatment effects. A bias of zero means that an estimator successfully removes all differences between the experimental control group and the non-experimental comparison group. We have arranged the estimators in the tables in logical groups. OLS heads up the table followed by the two cell matching estimators, followed by the basic cross-sectional estimators, followed by the bias-corrected matching estimators, followed by the longitudinal difference-in-differences matching estimators.

We can characterize the results in Table 7 in terms of five main patterns, four of which relate back to conclusions drawn in Heckman et al. (1997, 1998c) for JTPA viewed as a single treatment. First, we find little evidence of large biases from applying matching in this context, with these conditioning variables, to the individual treatment streams. We do not want to push this finding very hard as, given our standard errors, we can also not distinguish our estimates from a wide range of population bias values, both positive and negative. Moreover, we have substantively large point estimates for the biases in some cases, though less often for the better performing estimators. On the other hand, if the data wanted to send a strong signal that matching fails miserably here, they could have done so, but did not.

Second, our estimates reveal the possibility of some substantively meaningful cancellation in bias across treatment streams when aggregating JTPA into a single treatment. Third, the three simple estimators — OLS, matching on labor force status transition cells and propensity score stratification — tend to have lower variances than the other matching estimators. Of the three simple estimators, stratification on the propensity score clearly dominates. Indeed, its solid performance on all three dimensions comports with its frequent use in

Table 7 Adult males and females — bias estimates from propensity score matching

Estimator		Overall	CT-OS	OJT	Other
Males					
Outcome: sum of earnings over 18 months after RA/EL					
1. OLS	Bias	1323.0	258.7	1083.9	1943.0
	Std. err.[a]	1097.6	1826.6	1264.0	1639.3
	RMSE	1719.1	1844.9	1665.1	2542.2
2. LF transition cell matching	Bias	−1667.7	−3566.9	−2204.0	−1455.6
	Std. err.	1196.4	1640.0	1350.1	1575.3
	RMSE	2052.4	3925.9	2584.6	2144.8
3. P-score decile cell matching	Bias	49.6	−1834.9	−528.7	−2147.3
	Std. err.	1492.1	1683.8	1431.3	1612.7
	RMSE	1493.0	2490.4	1525.8	2685.5
4. 1 Nearest neighbor matching	Bias	−1281.6	512.2	−1332.5	−2969.9
	Std. err.	1979.2	−539.3	1500.5	2351.0
	RMSE	2357.9	743.8	2006.7	3787.8
5. Nearest 12 neighbors matching (optimal within 25 neighbors)[b]	Bias	176.4	−1044.3	−265.1	−716.5
	Std. err.	1849.4	−883.4	1381.6	2121.5
	RMSE	1857.8	1367.9	1406.8	2239.3
6. Optimal kernel[c]	Bias	−555.5	−551.5	−1238.2	−2589.4
	Std. err.	1149.5	2771.0	1803.6	1803.5
	RMSE	1276.6	2825.4	2187.8	3155.5
7. Optimal local linear[a]	Bias	−369.2	−2077.7	−1535.6	−865.0
	Std. err.	1325.8	2382.0	1495.7	2054.6
	RMSE	1376.3	3160.8	2143.6	2229.3
8. Bias-corrected 1 nearest neighbor matching	Bias	−968.3	−58.5	−1100.7	−2562.8
	Std. err.	1538.3	3633.2	1679.5	2313.8
	RMSE	1817.7	3633.6	2008.1	3452.8
9. Bias-corrected kernel matching	Bias	−36.4	−967.0	−149.9	−1237.0
	Std. err.	1008.1	2135.5	1534.1	1489.5
	RMSE	1008.8	2344.3	1541.4	1936.2
10. Bias-corrected local linear matching	Bias	−968.3	−677.7	−1100.7	−2562.8
	Std. err.	1538.3	2337.9	1679.5	2313.8
	RMSE	1817.7	2434.1	2008.1	3452.8
11. One nearest neighbor difference-in-differences matching	Bias	−1278.1	603.4	−2439.0	−1675.0
	Std. err.	1864.1	4695.0	3135.0	3143.0
	RMSE	2260.2	4733.6	3972.0	3561.4
12. Kernel difference-in-differences matching[c]	Bias	−344.5	−2157.6	717.4	2110.7
	Std. err.	1104.8	3185.8	1398.8	1744.9
	RMSE	1157.3	3847.7	1572.0	2738.5
Outcome: employment in quarter 6 after RA/EL					
1. OLS	Bias	0.135	0.165	0.093	0.165
	Std. err.[a]	0.048	0.085	0.053	0.061
	RMSE	0.143	0.186	0.107	0.176
2. LF transition cell matching	Bias	0.080	0.067	0.052	0.083
	Std. err.	0.050	0.078	0.054	0.059
	RMSE	0.094	0.102	0.075	0.101

Table 7 continued

Estimator		Overall	CT-OS	OJT	Other
3. P-score decile cell matching	Bias	0.065	0.107	−0.006	0.038
	Std. err.	0.060	0.083	0.051	0.057
	RMSE	0.088	0.135	0.051	0.069
4. 1 Nearest neighbor matching	Bias	0.041	0.163	0.005	−0.019
	Std. err.	0.168	0.095	0.151	0.190
	RMSE	0.173	0.189	0.151	0.191
5. Nearest 12 neighbors matching (optimal within 25 neighbors)[b]	Bias	0.134	0.105	0.071	0.123
	Std. err.	0.155	0.088	0.137	0.175
	RMSE	0.205	0.137	0.154	0.214
6. Optimal kernel[d]	Bias	0.006	0.072	−0.049	−0.019
	Std. err.	0.063	0.138	0.068	0.052
	RMSE	0.063	0.155	0.084	0.055
7. Optimal local linear[d]	Bias	0.043	0.083	0.016	0.045
	Std. err.	0.281	0.334	0.427	0.262
	RMSE	0.284	0.344	0.427	0.266
8. Bias-corrected 1 nearest neighbor matching	Bias	0.043	0.125	0.005	−0.001
	Std. err.	0.070	0.157	0.082	0.112
	RMSE	0.082	0.201	0.082	0.112
9. Bias-corrected kernel matching	Bias	0.029	0.076	−0.006	0.090
	Std. err.	0.061	0.129	0.077	0.078
	RMSE	0.068	0.149	0.078	0.119
10. Bias-corrected local linear matching	Bias	0.051	0.056	0.035	0.089
	Std. err.	0.067	0.134	0.081	0.081
	RMSE	0.084	0.145	0.089	0.120
11. One nearest neighbor difference-in-differences matching	Bias	−0.086	−0.021	−0.081	−0.087
	Std. Err.	0.191	0.379	0.245	0.204
	RMSE	0.209	0.379	0.258	0.222
12. Kernel difference-in-differences matching[d]	Bias	0.074	0.008	0.013	0.057
	Std. err.	0.057	0.141	0.074	0.085
	RMSE	0.093	0.142	0.075	0.102
Females					
Outcome: sum of earnings over 18 months after RA/EL					
1. OLS	Bias	1321.4	1031.4	2052.4	695.8
	Std. err.[a]	483.7	609.2	693.7	730.7
	RMSE	1407.2	1197.9	2166.5	1009.0
2. LF transition cell matching	Bias	1569.6	849.7	2291.7	1451.3
	Std. err.	445.6	539.7	640.6	672.7
	RMSE	1631.6	1006.5	2379.6	1599.6
3. P-score decile cell matching	Bias	1181.6	794.6	1556.6	127.6
	Std. err.	537.6	614.3	732.2	786.7
	RMSE	1298.1	1004.4	1720.2	797.0
4. 1 Nearest neighbor matching	Bias	949.1	1181.0	1797.4	−511.0
	Std. err.	936.2	995.6	1233.3	1885.3
	RMSE	1333.2	1544.7	2179.8	1953.3

Table 7 continued

Estimator			Overall	CT-OS	OJT	Other
5.	Nearest 18 neighbors matching (optimal within 20 neighbors)[b]	Bias	1062.0	796.2	1106.5	266.3
		Std. err.	639.0	663.6	848.8	800.9
		RMSE	1239.4	1036.5	1394.5	844.0
6.	Optimal kernel[e]	Bias	1176.1	393.2	1175.6	−564.3
		Std. err.	577.4	663.3	986.5	1145.0
		RMSE	1310.2	771.1	1534.7	1276.6
7.	Optimal local linear[e]	Bias	857.2	868.7	1380.6	−200.6
		Std. Err.	574.3	646.4	910.2	1303.2
		RMSE	1031.8	1082.8	1653.7	1318.5
8.	Bias-corrected 1 nearest neighbor matching	Bias	1229.9	1168.4	1576.8	−92.5
		Std. err.	723.6	908.0	1087.4	1435.3
		RMSE	1427.0	1479.8	1915.4	1438.2
9.	Bias-corrected kernel matching	Bias	1203.1	816.9	1052.8	−551.9
		Std. err.	517.0	641.6	918.5	1164.7
		RMSE	1309.5	1038.7	1397.1	1288.8
10.	Bias-corrected local linear matching	Bias	1229.9	1168.4	1576.8	−92.5
		Std. err.	648.2	942.3	1032.8	1533.9
		RMSE	1390.2	1501.0	1884.9	1536.7
11.	One nearest neighbor difference-in-differences matching	Bias	1352.6	1862.4	1822.0	5.7
		Std. err.	820.8	1075.2	1561.2	1638.6
		RMSE	1582.1	2150.5	2399.4	1638.6
12.	Kernel difference-in-differences matching[e]	Bias	1292.0	754.1	2036.5	832.2
		Std. err.	415.7	605.3	674.4	776.9
		RMSE	1357.2	967.0	2145.3	1138.5
Outcome: employment in quarter 6 after RA/EL						
1.	OLS	Bias	0.089	0.083	0.111	0.070
		Std. err.[a]	0.031	0.042	0.040	0.046
		RMSE	0.094	0.093	0.118	0.083
2.	LF transition cell matching	Bias	0.093	0.056	0.137	0.085
		Std. err.	0.032	0.041	0.041	0.046
		RMSE	0.098	0.069	0.143	0.097
3.	P-score decile cell matching	Bias	0.088	0.068	0.114	0.012
		Std. err.	0.037	0.047	0.052	0.047
		RMSE	0.095	0.082	0.125	0.048
4.	1 Nearest neighbor matching	Bias	0.087	0.097	0.091	0.003
		Std. err.	0.050	0.069	0.084	0.090
		RMSE	0.100	0.119	0.124	0.090
5.	Nearest 18 neighbors matching (optimal within 20 neighbors)[b]	Bias	0.075	0.071	0.103	0.065
		Std. err.	0.034	0.042	0.054	0.071
		RMSE	0.082	0.082	0.117	0.096
6.	Optimal kernel[f]	Bias	0.077	0.000	0.121	0.025
		Std. err.	0.030	0.058	0.043	0.054
		RMSE	0.082	0.058	0.129	0.060

Table 7 continued

Estimator			Overall	CT-OS	OJT	Other
7.	Optimal local linear[f]	Bias	0.074	0.024	0.097	0.034
		Std. err.	0.044	0.048	0.064	0.067
		RMSE	0.086	0.054	0.116	0.075
8.	Bias-corrected	Bias	0.098	0.083	0.089	0.031
	1 nearest neighbor matching	Std. err.	0.047	0.066	0.075	0.070
		RMSE	0.109	0.106	0.116	0.077
9.	Bias-corrected	Bias	0.077	0.052	0.110	0.055
	kernel matching	Std. err.	0.029	0.057	0.039	0.053
		RMSE	0.083	0.077	0.116	0.076
10.	Bias-corrected	Bias	0.098	0.083	0.089	0.031
	local linear matching[f]	Std. err.	0.039	0.067	0.072	0.085
		RMSE	0.106	0.107	0.115	0.090
11.	One nearest neighbor	Bias	0.057	0.075	0.128	0.066
	difference-in-differences matching	Std. err.	0.060	0.101	0.097	0.093
		RMSE	0.083	0.126	0.161	0.114
12.	Kernel	Bias	0.072	0.028	0.137	0.068
	difference-in-differences matching[f]	Std. err.	0.055	0.069	0.052	0.061
		RMSE	0.090	0.075	0.147	0.091

[a] Robust standard errors for Estimator 1 (OLS) and bootstrap standard errors (50 repetitions) for estimators 4 to 13

[b] Using 12(18) neighbors in nearest neighbor matching minimizes RMSE among the first 25 neighbors for males (females)

[c] Optimal kernel and bandwidth are chosen with cross-validation to minimize RMSE
Estimator 6 (Optimal Kernel) Epanechnikov 0.0140 (Epanechnikov 0.0062 in CT-OS)
Estimator 7 (Optimal Local Linear) Tricube 0.2962 (Epanechnikov 0.0985 in CT-OS)
Estimator 12 (Kernel D-I-D matching) Gaussian 0.0273 (Epanechnikov 0.0058 in CT-OS)

[d] Optimal kernel and bandwidth are chosen with cross-validation to minimize RMSE
Estimator 6 (Optimal Kernel) Gaussian 0.0518 (Epanechnikov 0.0066 in CT-OS)
Estimator 7 (Optimal Local Linear) Gaussian 0.1344 (Epanechnikov 0.0107 in CT-OS)
Estimator 12 (Kernel D-I-D matching) Epanechnikov 0.0570 (Epanechnikov 0.0057 in CT-OS)

[e] Optimal kernel and bandwidth are chosen with cross-validation to minimize RMSE
Estimator 6 (Optimal Kernel) Gaussian 0.0045
Estimator 7 (Optimal Local Linear) Tricube 0.059
Estimator 12 (Kernel D-I-D matching) Gaussian 0.147

[f] Optimal kernel and bandwidth are chosen with cross-validation to minimize RMSE
Estimator 6 (Optimal Kernel) Tricube 0.0137
Estimator 7 (Optimal Local Linear) Tricube 0.123
Estimator 12 (Kernel D-I-D matching) Tricube 0.034

the applied statistics literature and suggests its value as a baseline for more complicated matching schemes. The other two estimators do less well in terms of bias, leading to relatively mediocre RMSEs.

Fourth, Heckman et al. (1997, 1998a) argued that, in general, the details of the matching method do not matter much. Our results suggest a more nuanced picture, keeping in mind, as always, the imprecision both in the bias estimates and

in the variance estimates (and thereby in the RMSE estimates). In particular, we find that single nearest neighbor matching performs quite poorly, consistent with its performance in the very useful Monte Carlo analysis in Frölich (2004). This suggests the wisdom of the general preference for kernel matching in the applied economics literature. Bias corrected single nearest neighbor matching often does better in terms of both bias and variance, supporting the use of ex post regression following matching in the applied statistics literature. We do not observe consistent improvements in RMSE from ex post regression for the other cross-sectional matching estimators in our data. Also, nearest neighbor with a number of neighbors (sometimes surprisingly large) chosen by cross-validation generally yields a noticeably lower variance than single nearest neighbor matching, as one would expect, but only modestly higher bias, so that in RMSE terms it generally wins the contest between the two estimators.

Fifth and finally, as noted in Heckman and Smith (1999), no strong pattern emerges in terms of biases, variances or RMSEs that would imply a clear choice between cross-sectional and difference-in-differences matching in this context. This result differs strongly from that found by Smith and Todd (2005a) using the Supported Work data. This difference arises from the fact that the NJS data, unlike the Supported Work data, do not embody time invariant biases resulting from geographic mismatch or from the use of outcome variables measured differently for the treated and untreated units.

6 Conclusions

Multi-treatment programs appear in many contexts, in particular that of active labor market policy. In this paper, we have considered the trade-offs involved in evaluating such programs as disaggregated treatments rather than an aggregate whole, and have illustrated some of our points using data from the U.S. National JTPA Study. Though our evidence suffers from the relatively small sample sizes that remain once we disaggregate, we nonetheless find interesting differences in experimental estimates and in the determinants of participation across the three treatment streams. These differences add to our understanding of the program and illustrate the potential for cancellation across treatments when aggregating to hide relevant differences among them. We also add to the literature on the performance of alternative matching estimators, where we have more to say than did the aggregative analyses in Heckman et al. (1997, 1998a). In particular, our results highlight the relatively poor performance of the widely used single nearest neighbor matching estimator.

Acknowledgments We gratefully acknowledge financial support from the Social Sciences and Humanities Research Council of Canada and the CIBC Chair in Human Capital and Productivity at the University of Western Ontario. We thank Chris Mitchell for his excellent research assistance, Michael Lechner for helpful discussions, two anonymous referees and our editor, Bernd Fitzenberger, for their helpful comments and Dan Black for his Stata cross-validation code.

References

Abadie A, Imbens G (2006) On the failure of the bootstrap for matching estimators. Unpublished manuscript, University of California at Berkeley

Andrews D, Buchinsky M (2000) A three-step method for choosing the number of bootstrap repetitions. Econometrica 68:23–51

Angrist J, Krueger K (1999) Empirical strategies in labor economics. In: Ashenfelter O, Card D (eds) Handbook of Labor Economics Vol 3A. North-Holland, Amsterdam, pp 1277–1366

Ashenfelter O (1978) Estimating the effect of training programs on earnings. Rev Econ Stat 6:47–57

Black D, Smith J (2004) How robust is the evidence on the effects of college quality? Evidence from matching. J Econ 121:99–124

Black D, Smith J, Berger M, Noel B (2003) Is the threat of reemployment services more effective than the services themselves? Evidence from the UI system using random assignment. Am Econ Rev 93:1313–1327

Bloom H, Orr L, Cave G, Bell S, Doolittle F (1993) The National JTPA Study: title II-A impacts on earnings and employment at 18 Months. Abt Associates, Bethesda

Bloom H, Orr L, Bell S, Cave G, Doolittle F, Lin W, Bos J (1997) The benefits and costs of JTPA title II-A programs: key findings from the National Job Training Partnership Act study. J Hum Resources 32:549–576

Card D, Sullivan D (1988) Measuring the effect of subsidized training programs on movements in and out of employment. Econometrica 56:497–530

Courty P, Marschke G (2004) An empirical investigation of gaming responses to explicit performance incentives. J Labor Econ 22:23–56

Dehejia R, Wahba S (1999) Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. J Am Stat Assoc 94:1053–1062

Dehejia R, Wahba S (2002) Propensity score matching methods for non-experimental causal studies. Rev Econ Stat 84:139–150

Devine T, Heckman J (1996) The consequences of eligibility rules for a social program: a study of the Job Training Partnership Act. Res Labor Econ 15:111–170

Dolton P, Smith J, Azevedo JP (2006) The econometric evaluation of the new deal for lone parents. Unpublished manuscript, University of Michigan

Doolittle F, Traeger L (1990) Implementing the National JTPA Study. Manpower Demonstration Research Corporation, New York

Dorset R (2006) The New Deal for Young People: effect on the labor market status of young men. Labour Econ 13:405–422

Fan J, Gijbels I (1996) Local polynomial modeling and its applications. Chapman and Hall, New York

Fisher R (1935) The design of experiments. Oliver and Boyd, London

Fitzenberger B, Speckesser S (2005) Employment effects of the provision of specific professional skills and techniques in Germany. IZA Working paper no. 1868

Frölich M (2004) Finite sample properties of propensity score matching and weighting estimators. Rev Econ Stat 86:77–90

Frölich, M (2006) A note on parametric and nonparametric regression in the presence of endogenous control variables. IZA working paper no. 2126

Galdo J, Smith J, Black D (2006) Bandwidth selection and the estimation of treatment effects with nonexperimental data. Unpublished manuscript, University of Michigan

Gerfin M, Lechner M (2002) Microeconometric evaluation of active labour market policy in Switzerland. Econ J 112:854–803

Heckman J (1979) Sample selection bias as a specification error. Econometrica 47:153–161

Heckman J, Hotz VJ (1989) Choosing among alternative nonexperimental methods for estimating the impact of training programs. J Am Stat Assoc 84:862–874

Heckman J, Navarro S (2004) Using matching, instrumental variables, and control functions to estimate economic choice models. Rev Econ Stat 86:30–57

Heckman J, Smith J (1999) The pre-programme earnings dip and the determinants of participation in a social programme: implications for simple program evaluation strategies. Econ J 109:313–348

Heckman J, Smith J (2000) The sensitivity of experimental impact estimates: evidence from the National JTPA Study. In: Blanchflower D, Freeman R (eds) Youth employment and joblessness in advanced countries. University of Chicago Press, Chicago
Heckman J, Smith J (2004) The determinants of participation in a social program: evidence from a prototypical job training program. J Labor Econ 22:243–298
Heckman J, Todd P (1995) Adapting propensity score matching and selection models to choice-based samples. Unpublished manuscript, University of Chicago
Heckman J, Ichimura H, Todd P (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training program. Rev Econ Stud 64:605–654
Heckman J, Ichimura H, Smith J, Todd P (1998a) Characterizing selection bias using experimental data. Econometrica 66:1017–1098
Heckman J, Lochner L, Taber C (1998b) Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. Rev Econ Dynam 1:1–58
Heckman J, Smith J, Taber C (1998c) Accounting for dropouts in evaluations of social programs. Rev Econ Stat 80:1–14
Heckman J, LaLonde R, Smith J (1999) The economics and econometrics of active labor market programs. In: Ashenfelter O, Card D (eds) Handbook of Labor Economics, Vol 3A. North-Holland, Amsterdam, pp 1865–2097
Heckman J, Hohmann N, Smith J, Khoo M (2000) Substitution and dropout bias in social experiments: a study of an influential social experiment. Q J Econ 115:651–694
Heckman J, Heinrich C, Smith J (2002) The performance of performance standards. J Hum Resources 36:778–811
Heinrich C, Marschke G, Zhang A (1999) Using administrative data to estimate the cost-effectiveness of social program services. Unpublsihed manuscript, Univerity of Chicago
Ho D, Kosuke I, King G, Stuart E (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Forthcoming in: Political Analysis
Imbens G (2000) The role of the propensity score in estimating dose-response functions. Biometrika 87:706–710
Kordas G, Lehrer S (2004) Matching using semiparametric propensity scores. Unpublished manuscript, Queen's University
Kemple J, Doolittle F, Wallace J (1993) The National JTPA Study: site characteristics and participation patterns. Manpower Demonstration Research Corporation, New York
LaLonde R (1986) Evaluating the econometric evaluations of training programs using experimental data. Am Econ Rev 76:604–620
Lechner M (2001) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Lechner M, Pfeiffer P (eds) Econometric evaluation of labour market policies. Physica, Heidelberg
Lechner M, Smith J (2007) What is the value added by caseworkers? Labour Econ 14:135–151
Lechner M, Miquel R, Wunsch C (2008) The curse and blessing of training the unemployed in a changing economy: the case of East Germany after unification. Forthcoming in: German Economic Review
Lise J, Seitz S, Smith J (2005) Equilibrium policy experiments and the evaluation of social programs. NBER working paper no. 10283
Manski C (1996) Learning about treatment effects from experiments with random assignment to treatment. J Hum Resources 31:707–733
Michalopolous C, Tattrie D, Miller C, Robins P, Morris P, Gyarmati D, Redcross C, Foley K, Ford R (2002) Making work pay: final report on the Self-Sufficiency Project for long-term welfare recipients. Social Research and Demonstration Corporation, Ottawa
Neyman J (1923) Statistical problems in agricultural experiments. J R Stat Soc 2:107–180
Orr L, Bloom H, Bell S, Lin W, Cave G, Doolittle F (1994) The National JTPA Study: impacts, benefits and costs of title II-A. Abt Associates, Bethesda
Pagan A, Ullah A (1999) Nonparametric econometrics. Cambridge University Press, Cambridge
Pechman J, Timpane M (1975) Work incentives and income guarantees: the New Jersey negative income tax experiment. Brookings Institution, Washington DC
Plesca M (2006) A general equilibrium evaluation of the employment service. Unpublished manuscript, University of Guelph

Quandt R (1972) Methods of estimating switching regressions. J Am Stat Assoc 67:306–310
Racine J, Li Q (2005) Nonparametric estimation of regression functions with both categorical and continuous data. J Econ 119:99–130
Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55
Rosenbaum P, Rubin D (1984) Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 79:516–524
Rosenbaum P, Rubin D (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 39:33–38
Roy AD (1951) Some thoughts on the distribution of earnings. Oxford Econ Pap 3:135–146
Rubin D (1974) Estimating causal effects of treatments in randomized and non-randomized studies. J Educ Psychol 66:688–701
Smith J, Todd P (2005a) Does matching overcome LaLonde's critique of nonexperimental methods? J Econ 125:305–53
Smith J, Todd P (2005b) Rejoinder J Econ 125:365–375
Smith J, Whalley A (2006) How well do we measure public job training? Unpublished manuscript, University of Michigan
Zhao Z (2004) Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. Rev Econ Stat 86:91–107

Appendix

Table A1 Variable definitions

OUTCOMES
Sum of quarterly earnings in the first 6 quarters after RA/EL (RA/EL is the date of random assignment for the controls and the date of eligibility screening for the ENPs.). Constructed from the average monthly earnings per quarter variable used in Heckman et al. (1997, 1998c)
Indicator of positive earnings in the sixth quarter after RA/EL
BACKGROUND VARIABLES FROM THE LONG BASELINE SURVEY
Age
Indicators for ages 30–39, 40–49 and 50–54
Education
Highest grade of formal schooling that the respondent had completed as of the long baseline interview. Recoded into the following indicator variables for particular ranges of the highest grade completed: highest grade completed < 10 at interview time, between 10–11, 12, 13–15, and >15
Marital status
The respondent's marital status during the 12 months prior to RA/EL: currently married at RA/EL, last married 1–12 months prior to RA/EL, last married >12 months prior to RA/EL, and single or never married at RA/EL. Only the first category, currently married at RA/EL, was used in the baseline specification
Family earnings in the year prior to the baseline interview
The sum of the total earnings of all related household members, including the respondent, in the year prior to the baseline interview. It includes only persons in the household at the time of the interview. The total is set to missing if the employment status is missing for any related household member or if the annual earnings are missing for any employed related household member. The indicators for individual categories, which include imputations, are: family income between \$0 and \$3,000, between \$3,001 and \$9,000, between \$9,001 and \$15,000, and greater than \$15,000.
Quarterly welfare pattern before RA/EL
Pattern of quarterly welfare receipt in the two quarters up to and including RA/EL. The quarterly welfare receipt variables are set to 1 if the respondent received AFDC, food stamps, or general

Table A1 continued

assistance in any of the months in the quarter and to zero if they received none of these in all of the months of the quarter. The coding of the patterns created by the two quarterly variables, and the corresponding indicator variables, are as follows:
no welfare → no welfare, no welfare → welfare, welfare → no welfare and welfare → welfare.

Two most recent labor force status values before RA/EL
Two most recent values of the monthly labor force status in the 7 months up to and including the month of RA/EL. The values of this variable, and the corresponding indicator variables, are: emp → emp, unm → emp, olf → emp, emp → unm, unm → unm, olf → unm, emp → olf, unm → olf and olf → olf

Quarterly earnings in the six most recent quarters prior to RA/EL
Average earnings per quarter are constructed from monthly measures self-reported by individuals

IMPUTATIONS
Missing values due to item non-response were imputed for the variables listed earlier. Missing values for continuous variables, such as household members, were imputed using the predicted values from a linear regression. Missing values of dichotomous variables, such as the presence of any own children in the household, were replaced with the predicted probabilities estimated in a logit equation. Missing values of indicator variables corresponding to particular values of categorical variables with more than two categories, such as the five indicators for the categories of the highest grade completed variable, were replaced by the predicted probabilities obtained from a multinomial logit model with the categorical variable as the dependent variable. In all cases, the estimating equations used to produce the imputations included: indicators for race/ethnicity, indicators for age categories, indicators for receipt of a high school diploma or a GED, and site indicators. All variables were interacted with a control group indicator. Variables included were chosen because they had no missing values in the sample. Separate imputation models were estimated for adult males and adult females

Table A2 Optimal kernel and bandwidth

Kernel type	Gaussian	Epanechnikov	Tricube
Cross-validation analysis for kernel matching			
Earnings outcome			
Adult males			
Optimal bandwidth	0.0116	**0.0140**	0.0140
Smallest RMSE	4738.5560	**4724.0550**	4742.8010
Minimum bandwidth 0.004			
Maximum bandwidth 0.635			
Grid size 55			
Sample size (ENP) 357			
Adult females			
Optimal bandwidth	**0.0045**	0.0097	0.0106
Smallest RMSE	**3043.9586**	3044.8603	3044.5481
Minimum bandwidth 0.001			
Maximum bandwidth 0.271			
Grid size 64			
Sample size (ENP) 870			
Employment outcome			
Adult males			
Optimal bandwidth	**0.0518**	0.1222	0.1478
Smallest RMSE	**0.4287**	0.4292	0.4293
Minimum bandwidth 0.004			
Maximum bandwidth 0.618			
Grid size 55			
Sample size (ENP) 367			

Table A2 continued

Kernel type	Gaussian	Epanechnikov	Tricube
Adult females			
Optimal bandwidth	0.2898	0.0125	**0.0137**
Smallest RMSE	0.5006	0.4992	**0.4992**
Minimum bandwidth 0.001			
Maximum bandwidth 0.290			
Grid size 65			
Sample size (ENP) 896			
Cross-validation analysis for local linear matching			
Earnings outcome			
Adult males			
Optimal bandwidth	0.0645	0.0586	**0.2962**
Smallest RMSE	4682.6660	4669.4770	**4594.7070**
Minimum bandwidth 0.004			
Maximum bandwidth 0.635			
Grid size 55			
Sample size (ENP) 357			
Adult females			
Optimal bandwidth	0.2714	0.2714	**0.0591**
Smallest RMSE	3054.8267	3056.1162	**3012.5712**
Minimum bandwidth 0.001			
Maximum bandwidth 0.271			
Grid size 64			
Sample size (ENP) 860			
Employment outcome			
Adult males			
Optimal bandwidth	**0.1344**	0.4640	0.5104
Smallest RMSE	**0.4238**	0.4238	0.4250
Minimum bandwidth 0.004			
Maximum bandwidth 0.618			
Grid size 55			
Sample size (ENP) 367			
Adult females			
Optimal bandwidth	0.2898	0.0221	**0.1229**
Smallest RMSE	0.5014	0.4986	**0.4957**
Minimum bandwidth 0.001			
Maximum bandwidth 0.290			
Grid size 65			
Sample size (ENP) 896			
Cross-validation analysis for difference-in-differences kernel matching			
Earnings outcome			
Adult males			
Optimal bandwidth	**0.0273**	0.0645	0.0709
Smallest RMSE	**3834.1230**	3840.9380	3841.2520
Minimum bandwidth 0.004			
Maximum bandwidth 0.635			
Grid size 55			
Sample size (ENP) 357			
Adult females			
Optimal bandwidth	**0.1472**	0.2868	0.2868
Smallest RMSE	**2271.1581**	2271.0654	2272.0574
Minimum bandwidth 0.001			
Maximum bandwidth 0.287			
Grid size 64			
Sample size (ENP) 823			

Table A2 continued

Kernel type	Gaussian	Epanechnikov	Tricube
Employment outcome			
Adult males			
Optimal bandwidth	0.0389	**0.0570**	0.0570
Smallest RMSE	0.4870	**0.4849**	0.4855
Minimum bandwidth 0.004			
Maximum bandwidth 0.618			
Grid size 55			
Sample size (ENP) 367			
Adult females			
Optimal bandwidth	0.1708	0.0210	**0.0338**
Smallest RMSE	0.5218	0.5210	**0.5209**
Minimum bandwidth 0.001			
Maximum bandwidth 0.275			
Grid size 64			
Sample size (ENP) 858			

1 The endpoints of the grid for bandwidth search are (Xmax-Xmin)/N and (Xmax-Xmin)/2. Each step increments the previous bandwidth by a factor of 1.1

2 Within each demographic group we use the same comparison group of ENPs for all three treatment streams; as a result, the optimal bandwidth is the same as well. The exception is adult males in the CT-OS treatment stream, for whom we adopt a slightly different propensity score specification due to the small sample size

Table A3 Balancing tests for adult males and females

	Overall	CT-OS	OJT	Other
Adult males				
Nearest neighbor standardized differences				
Site: Fort Wayne	−11.65	0.00	1.00	3.94
Site: Jersey City	−5.91	0.00	−10.74	−11.24
Site: providence	16.87	0.00	6.64	4.73
Race: black	5.40	6.16	−13.56	0.00
Race: other	−5.58	−19.97	−6.37	−8.45
Age	6.57	17.11	−0.83	−7.63
Age squared	6.92	14.48	−2.38	−5.63
Educ. <10 years	−3.30	2.42	−12.50	15.81
Educ. 10–11years	1.39	−9.67	5.72	−7.31
Married at RA/EL	−2.65	−9.05	−5.15	−8.12
Fam. Inc. 3 K–9 K	6.48	.	−17.87	15.08
Fam. Inc. 9 K–15 K	8.97	5.73	−1.90	16.88
Fam. Inc. >15 K	−8.08	−17.97	−6.22	−1.36
LF into unempl.	−2.86	26.61	−6.57	7.87
LF into OLF	0.64	−14.26	3.26	6.34
Earnings Q-1	−4.63	4.63	−6.67	−6.63
Earnings Q-2	4.04	1.77	−8.80	0.88
Earnings Q-3 to Q-6	0.05	−17.55	−3.79	−14.66
Standardized differences summary for cross-sectional estimators (estimators 4 to 7 from Table 7)				
Nearest neighbor				
Maximum absolute standardized difference	16.87	26.61	17.87	16.88
Instances when absolute std.dif. >20	0	1	0	0
Average absolute standardized difference	5.67	9.85	6.66	7.92

Table A3 continued

	Overall	CT-OS	OJT	Other
Optimal nearest neighbors (12)				
Maximum absolute standardized difference	17.62	38.79	14.59	29.84
Instances when absolute std.dif. >20	0	6	0	1
Average absolute standardized difference	5.14	13.58	6.14	12.28
Optimal kernel (Epanechnikov 0.014)				
Maximum absolute standardized difference	21.17	37.46	23.02	31.39
Instances when absolute std.dif. >20	1	11	1	1
Average absolute standardized difference	7.52	20.19	8.34	10.91
Optimal local linear (Gaussian 0.0045)				
Maximum absolute standardized difference	10.16	21.83	14.25	23.61
Instances when absolute std.dif. > 20	0	2	0	1
Average absolute standardized difference	3.99	9.60	6.15	8.95
Adult females				
Nearest neighbor standardized differences				
Site: Fort Wayne	−0.65	5.04	−14.50	−3.58
Site: Jersey City	−4.71	−7.86	5.44	8.89
Site: providence	−7.02	1.21	−1.19	−5.74
Race: black	−11.70	−3.38	1.88	−16.95
Race: other	7.18	−0.98	0.85	−1.23
Age	0.57	1.86	−2.97	−0.94
Age squared	0.52	2.26	−2.87	0.73
HS dropout	8.90	4.83	3.09	−7.36
Educ. >13years	−10.62	9.33	4.90	14.79
Married at RA/EL	7.23	−4.70	−5.08	−1.07
No welf.→welf.	10.78	−18.25	0.00	−14.04
Welf.→no welf.	−11.41	9.97	−9.40	−21.35
Welf.→welf.	−1.59	12.00	12.62	3.51
Welfare NA	1.16	−1.88	−1.42	6.25
Fam. Inc. 3 K–9 K	5.49	0.58	17.95	−8.79
Fam. Inc. 9 K–15 K	−6.85	−12.90	2.97	−10.66
Fam. Inc. >15 K	7.19	−4.05	−0.32	1.79
LF emp→emp	2.66	5.39	10.46	1.62
LF emp→olf	1.51	−0.90	−8.58	8.86
LF into unempl.	1.56	17.67	7.12	11.07
LF olf→emp	−0.53	−8.76	−12.37	−4.87
LF olf→unm	1.15	−0.42	−9.02	6.58
LF olf→olf	1.91	−10.13	−2.62	−6.59
Standardized differences summary for cross-sectional estimators (estimators 4 to 7 from Table 7)				
Nearest neighbor				
Maximum absolute standardized difference	11.70	18.25	17.95	21.35
Instances when absolute std.dif. >20	0	0	0	1
Average absolute standardized difference	4.91	6.28	5.98	7.27
Optimal nearest neighbors (18)				
Maximum absolute standardized difference	11.16	6.90	21.98	15.90
Instances when absolute std.dif. >20	0	0	2	0
Average absolute standardized difference	4.31	2.95	5.98	5.27
Optimal kernel (Tricube 0.2962)				
Maximum absolute standardized difference	11.30	15.21	10.52	19.39
Instances when absolute std.dif. >20	0	0	0	0
Average absolute standardized difference	3.79	6.71	4.80	7.11
Optimal local linear (Tricube 0.0591)				
Maximum absolute standardized difference	13.88	8.08	17.22	16.13
Instances when absolute std.dif. >20	0	0	0	0
Average absolute standardized difference	3.78	2.66	4.71	5.24