DISCUSSION PAPER SERIES

IZA DP No. 4451

Evaluating Nonexperimental Estimators for Multiple Treatments: Evidence from Experimental Data

Carlos A. Flores
Oscar A. Mitnik

September 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Evaluating Nonexperimental Estimators for Multiple Treatments: Evidence from Experimental Data

Carlos A. Flores
University of Miami

Oscar A. Mitnik
University of Miami
and IZA

Discussion Paper No. 4451
September 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 4451
September 2009

ABSTRACT

Evaluating Nonexperimental Estimators for Multiple Treatments: Evidence from Experimental Data*

This paper assesses the effectiveness of unconfoundedness-based estimators of mean effects for multiple or multivalued treatments in eliminating biases arising from nonrandom treatment assignment. We evaluate these multiple treatment estimators by simultaneously equalizing average outcomes among several control groups from a randomized experiment. We study linear regression estimators as well as partial mean and weighting estimators based on the generalized propensity score (GPS). We also study the use of the GPS in assessing the comparability of individuals among the different treatment groups, and propose a strategy to determine the overlap or common support region that is less stringent than those previously used in the literature. Our results show that in the multiple treatment setting there may be treatment groups for which it is extremely difficult to find valid comparison groups, and that the GPS plays a significant role in identifying those groups. In such situations, the estimators we consider perform poorly. However, their performance improves considerably once attention is restricted to those treatment groups with adequate overlap quality, with difference-in-difference estimators performing the best. Our results suggest that unconfoundedness-based estimators are a valuable econometric tool for evaluating multiple treatments, as long as the overlap quality is satisfactory.

JEL Classification: C13, C14, C21

Keywords: multiple treatments, nonexperimental estimators, generalized propensity score

Corresponding author:

Oscar A. Mitnik
Department of Economics
University of Miami
P.O. Box 248126
Coral Gables, FL 33124-6550
USA
E-mail: omitnik@miami.edu

* We thank Chris Bollinger, Alfonso Flores-Lagunes, Laura Giuliano, Guido Imbens, Carlos Lamarche, Phil Robins and seminar participants at the 2009 SOLE meetings, 2009 AEA meetings, University of Kentucky, Second Joint IZA/IFAU Conference on Labor Market Policy Evaluation, and University of Miami Labor Discussion Group for useful comments and suggestions. Bryan Mueller and Yongmin Zang provided excellent research assistance.

1 Introduction

Nonexperimental methods are widely used in economics and other disciplines to evaluate government programs and policies, and many other types of interventions. In the absence of an experiment, these methods are in many situations the only alternative. Among them, those based on a selection-on-observables or unconfoundedness assumption play a very important role. This assumption states that selection into the intervention or "treatment" is random (i.e., exogenous) conditional on a set of observable covariates. Most of the research on these methods has been on estimation of average treatment effects of a binary treatment (i.e., individuals either participate in a program or not) on an outcome.[1] In practice, however, individuals are usually exposed to different doses of the treatment or to more than one treatment. As a result, there has been a growing interest in evaluating programs or interventions in which the treatment is multivalued or there are multiple treatments,[2] and in different methods to evaluate such treatments.[3] Unfortunately, very little is known about the effectiveness of these methods in terms of reducing the potential selection bias present in nonexperimental evaluations of multiple treatments.

This paper contributes to the literature by assessing the performance of econometric methods based on an unconfoundedness assumption in estimating average treatment effects for multiple or multivalued treatments. To our knowledge, this is the first paper to address this issue. We study linear regression estimators as well as partial mean and weighting estimators based on the generalized propensity score (GPS), defined as the probability of receiving a particular treatment (or treatment level) conditional on covariates. In addition, we analyze the role of the GPS in identifying, for every individual in the population, individuals that are comparable in terms of observable characteristics in each of the treatment groups. This is a key element in the estimation of population effects in a multiple treatment setting.

The classical approach in economics when using an unconfoundedness assumption to estimate treatment effects has been the use of linear regression to adjust for differences in the covariates. Most of the recent research on this area has been on developing more flexible ways to control for these differences, and has focused mainly on the binary treatment case. Many semiparametric estimators have been proposed (e.g., Hahn, 1998; Heckman et al., 1997, 1998b; Hirano et al., 2003; Abadie and Imbens, 2006), some of which have been shown to achieve the semiparametric efficiency bound derived by Hahn (1998). Rosenbaum and Rubin (1983) showed that in order to control for observed variables it is sufficient to control for the propensity score, defined as the probability of receiving treatment conditional on the covariates. The propensity score plays a critical role in many of the recently developed semiparametric estimators. It is also key in identifying observations for which it is difficult to find comparable individuals in the opposite treatment arm in terms of the

[1] For a review of this literature see, for instance, Heckman et al. (1999), Imbens (2004), and Imbens and Wooldridge (2009).

[2] See, for example, Lechner (2002a, 2002b), Behrman et al. (2004), Frölich et al. (2004), Kluve et al. (2007), Plesca and Smith (2007), Mitnik (2008), and Flores et al. (2009).

[3] See, for instance, Imbens (2000), Lechner (2001), Hirano and Imbens (2004), Imai and van Dyk (2004), Abadie (2005), Flores (2007) and Cattaneo (2009). For a survey of some of these methods see Frölich (2004b).

covariates, and in restricting attention to the usually called "overlap" or "common support" region (e.g., Heckman et al., 1997; Dehejia and Wahba, 1999, 2002).

Given the importance and prevalence of multivalued and multiple treatments in practice, more recently there has been a growing interest in extending to this setting some of the results from the binary treatment case. Imbens (2000) and Lechner (2001) generalized the main results in Rosenbaum and Rubin (1983) to the multivalued (or multiple) treatment case, and Hirano and Imbens (2004) further extended them to the continuous treatment case.[4] For the multivalued setting, Imbens (2000) proposed partial mean and weighting estimators based on the GPS, and Cattaneo (2009) derived the semiparametric efficiency bound and introduced estimators that attain it. We consider several of these GPS-based estimators in this paper.

Two approaches have been used in the literature to assess the value of methods based on the unconfoundedness assumption for estimation of binary treatment effects (Imbens, 2004). The first approach relies on data from a randomized experiment and nonexperimental control groups, for instance, from alternative data sets (e.g., Lalonde, 1986) or from different locations (e.g., Friedlander and Robins, 1995). Estimators based on the unconfoundedness assumption are then applied to the nonexperimental control group and the experimental treatment group and, to asses the performance of the methods, the results are compared against those from the experiment (which are unbiased estimates of the population average treatment effect). This approach can also be implemented by applying the estimators to the nonexperimental and experimental control groups, in which case the benchmark is a zero treatment effect. The second approach is based on Monte Carlo simulations, where the performance of alternative estimators is evaluated under different scenarios (e.g., Frölich, 2004a; Zhao, 2004; Busso et al., 2009a, 2009b). Imbens (2004) discusses how the approach based on nonexperimental controls is aimed at assessing the plausibility of the unconfoundedness assumption and the value of the methods based on it; while the simulation-based approach is more helpful in identifying which particular estimators perform better in a given setting.[5] In this paper we follow the approach based on nonexperimental controls, since our purpose is to assess the likely reliability of the methods based on the unconfoundedness assumption in a multiple treatment setting. Since the estimators we analyze are implemented using an actual data set, our results are also informative about the relative performance of these estimators in a realistic (although particular) setting.

Since the influential paper by Lalonde (1986), many studies have evaluated the performance of nonexperimental methods for estimation of average treatment effects in a binary setting.[6] This literature has advanced our understanding of nonexperimental evaluations by specifying conditions under which methods based on the unconfoundedness assumption are more likely to replicate the

[4] An alternative extension of the results in Rosenbaum and Rubin (1983) to the multivalued and continuous cases is proposed by Imai and van Dyk (2004).

[5] Both approaches have advantages and disadvantages. For instance, relative to the approach based on nonexperimental controls, the artificial environments constructed in a simulation study are rarely representative of the situations found in practice; however, in a simulation it is possible to understand how alternative estimators behave in different environments by varying the parameters of the data generation process.

[6] Among others, see Fraker and Maynard, (1987), Heckman and Hotz (1989), Friedlander and Robins (1995), Heckman et al. (1997), Heckman et al. (1998a), Dehejia and Wahba (1999, 2002), Michalopoulos et al. (2004), Smith and Todd (2005), Dehejia (2005), and Mueser et al. (2007).

results from a randomized experiment. One of the main conclusions is the importance of comparing "comparable" individuals. For instance, Heckman et al. (1997) and Heckman et al. (1998a) stress the importance of comparing treatment and control groups from the same local labor market to which the same questionnaire is administrated, as well as having data on detailed labor market histories. This literature has also highlighted the importance of the propensity score in identifying regions of the data where treatment and control units are comparable in terms of observed characteristics.

In this paper we use data from the National Evaluation of Welfare-to-Work Strategies (NEWWS), a social experiment conducted in the U.S. in the 1990s in which individuals in several locations were randomly assigned to a control group or to different training programs. We resort to the availability of control groups in different locations to evaluate the performance of several unconfoundedness-based estimators for multiple treatments. We use these estimators to adjust for observable characteristics in order to eliminate differences in average outcomes among all control groups.[7] This strategy is similar to that previously used in the binary treatment context by Friedlander and Robins (1995), Michalopoulos et al. (2004) and Hotz et al. (2005). The key difference in our approach is that, while their focus is on pairwise comparisons between controls in different locations, we focus on *simultaneously* comparing all control groups, which requires the use of nonexperimental methods for multiple treatments.

Although relying on an experiment is in principle not required to perform this analysis, using data from the NEWWS experiment has several advantages for our purposes. First, all the individuals used in our study are welfare recipients at the time of randomization regardless of their location, which helps reducing the heterogeneity across sites. Second, the survey instruments and the data gathered for all the individuals are the same, and the data available is extremely rich; it includes individual and family characteristics, as well as welfare use and labor market histories. Third, we use the experiment itself to develop benchmark measures to assess the nonexperimental results in the paper. However, comparing individuals across different geographic locations makes our exercise much more difficult because of (potential) differences across local labor markets. Considering the key role given in the literature to comparing individuals with different treatment status within the same local labor market (e.g., Friedlander and Robins, 1995; Heckman et al., 1997; Heckman et al., 1998a), it is therefore important to keep in mind that we impose a very high yardstick on the nonexperimental estimators we study.

Our paper shows that one of the main issues that makes estimation of average effects more challenging, when moving from a binary to a multiple treatment setting, is that the overlap requirements become more demanding. We highlight the crucial role played by the GPS in assessing the quality of the overlap in the distribution of observable characteristics of the different treatment groups, and propose a strategy to determine the overlap or common support region. This strategy is less stringent than those previously used in the multiple treatment literature (e.g., Lechner, 2002a,

[7] An ideal data set for an evaluation like ours would include several (at least three) nonexperimental control groups all belonging to the *same* local labor market. Unfortunately, such data is not available to the best of our knowledge.

2002b; Frölich et al., 2004), and is motivated by a procedure commonly used in the binary setting (e.g., Dehejia and Wahba, 1999, 2002). Our paper also illustrates how, in a multiple treatment setting, one is more likely to encounter some treatment groups that are not comparable to the rest. We discuss the importance of the GPS in identifying those groups.

We find that the estimators perform poorly –in equalizing average outcomes across all control groups– when implemented using control groups in locations with extremely poor overlap in their GPS distributions and with very different local economic conditions. However, their performance improves considerably when applied to control groups in locations where the overlap quality is better and the local economic conditions are relatively more similar. The difference-in-difference estimators perform the best and compare well to benchmark measures derived from experimental data. The superior performance of the difference-in-difference estimators is consistent with previous findings in the binary treatment literature (e.g., Heckman et al., 1997; Heckman et al., 1998a; Smith and Todd, 2005). The overall improvement in the performance of the estimators when comparing individuals in more similar labor markets implies that, when the treatment groups belong to the same local labor market, the estimators are likely to perform better.[8]

In sum, our results suggest that the nonexperimental estimators studied are a valuable econometric tool when evaluating multiple or multivalued treatments in the absence of an experiment. Nevertheless, they also highlight that it is key to carefully analyze the overlap quality in applications, since the overlap issues that arise in the implementation of these methods when the treatment is binary are magnified in the multiple treatment setting.

The paper is organized as follows. We formalize the study setup in the following section. In Section 3 we present the estimators considered in the paper, and in Section 4 we discuss the use of the GPS in determining the overlap or common support region. In Section 5 we describe the data. In Section 6 we use the GPS to assess the comparability of the different control groups and present the results from implementing the nonexperimental estimators. Section 7 concludes.

2 Study setup

We exploit data from an experiment conducted in several locations to assess the effectiveness of the unconfoundedness-based estimators discussed in the following section. Within each of these sites, individuals were randomly assigned either to a control group or to one of alternative treatment groups. Based on this data, we formalize the study setup based on the potential outcomes approach developed by Neyman (1923) and extended by Rubin (1974) to nonexperimental settings. Each unit i in our sample, $i = 1, 2, \ldots, N$, comes from one of k possible sites. Let $D_i \in \{1, 2, \ldots, k\}$ be an indicator of the location of individual i. We denote the potential outcomes by $Y_i(t_d, d)$, where t_d stands for the treatment and d for the site. Hence, $Y_i(t_d, d)$ is the outcome unit i would obtain if she

[8] Michalopoulos et al. (2004) evaluate the performance of nonexperimental estimators for *binary treatments* relying also on the NEWWS experiment. As compared to our study, they use a different sample and outcomes (see Section 5 for details). Their conclusions, which are for the binary setting, are more negative than ours regarding the performance of nonexperimental estimators. However, as in our case, they conclude that comparing groups in the same local markets can be key in improving the estimators' performance.

were located in site d and given treatment t_d. Two features of our potential outcomes notation are worth mentioning. First, we let the potential outcome $Y(t_d, d)$ depend on d. Although it may be difficult to think of the location as something we can manipulate (i.e., a "treatment" in Holland's, 1986, sense), it is convenient for our purposes as our goal is to equate average outcomes for controls across all sites. Second, we let t_d depend on d, as sites may not offer the same treatments. For all sites, a value of t of zero denotes the control treatment, in which individuals are prevented from receiving any program services.

In this paper we focus exclusively on the individuals in the control groups, so we use only the potential outcomes at zero, or $Y(0, d)$. By focusing on the control treatment we minimize treatment heterogeneity across sites, as training programs differed across sites in terms of implementation, particular services offered, administration, etc.[9]

The data we observe for each unit is (Y_i, D_i, X_i), with X_i a set of pre-treatment covariates, and $Y_i = Y(0, D_i)$. Our parameters of interest in this paper are

$$\beta_d = E[Y(0, d)], \text{ for } d = 1, 2, \ldots, k. \tag{1}$$

The object in (1) gives the average potential outcome under the control treatment in location d for someone randomly selected from the *entire* population (i.e., from any of the k sites). In cases where d represents different levels of the treatment, (1) is commonly called the dose-response function in the statistics literature. In principle, we could take the expectation of $Y(0, d)$ over any subset of the union of the populations in each site. For instance, we could focus on $E[Y(0, d) | D_i = f]$ or $E[Y(0, d) | D_i = \{f, g\}]$ for any $d, f, g \in \{1, \ldots, k\}$. By focusing on the (entire) population effects in (1) we avoid selecting a particular site as a reference group, and having our results depend on this choice. On the other hand, the estimation problem becomes more challenging because we need to find, for each individual in the entire population, comparable individuals in each of the sites. In contrast, if we focused on $E[Y(0, d) | D_i = f]$, for instance, we would need to find comparable individuals in each site only for those individuals in the $D_i = f$ group.[10]

As documented in Section 5, the distribution of observable characteristics in our data differs systematically across the control groups in the k sites, so the controls from any particular location are not representative of the entire population. The main goal of this paper is to study whether the nonexperimental estimators described in the following section can properly adjust for the differences in these characteristics and equalize average outcomes for control individuals across all sites. Hence, the hypothesis we test is

$$\beta_1 = \beta_2 = \ldots = \beta_k. \tag{2}$$

The equalities in (2) form the basis of our analysis, as they imply that any of the k control groups

[9] Hotz et al. (2005) explicitly study program heterogeneity across sites. Their intuition is that if one is able to adjust for control group outcomes across sites, the comparison of adjusted outcomes for nominally equal treatments across sites may be interpreted as the effect of program heterogeneity across sites.

[10] Note that since the expectation in (1) is calculated over the entire population, any pairwise (population) average treatment effect can be calculated as $\beta_d - \beta_s$ for $d, s \in \{1, 2, \ldots, k\}$. For an analysis of estimation of pairwise differences of the form $E[Y(0, d) - Y(0, f) | D_i = \{d, f\}]$ for any $d, f \in \{1, \ldots, k\}$ see, for instance, Lechner (2001).

can be used to construct a valid counterfactual for the population average potential outcome in any of the other sites. As previously mentioned, the key difference between our approach and that in the existing literature (e.g., Michalopoulos et al., 2004; Hotz et al., 2005) is that we compare all locations *simultaneously*, which requires the use of nonexperimental methods for multiple treatments.

We assess the performance of the estimators presented in the following section in several ways. First, given estimates $\widehat{\beta}_1, \ldots, \widehat{\beta}_k$ of the corresponding parameters, we perform a Wald test of hypothesis (2). One drawback of this strategy is that it may be too sensitive to the variance of the estimators, in the sense that we could fail to reject the null hypothesis in (2) only because the variance of an estimator is high, and not necessarily because all the estimated $\widehat{\beta}$'s are sufficiently close to each other. Hence, a second approach is to directly compare overall measures of distance among the estimated means. Letting $\overline{\beta} = k^{-1} \sum_{d=1}^{k} \widehat{\beta}_d$, we define the following three distance measures: the root mean square distance ($rmsd$),

$$rmsd = \sqrt{\frac{1}{k} \sum_{d=1}^{k} \left(\widehat{\beta}_d - \overline{\beta} \right)^2}; \tag{3}$$

the mean absolute distance (mad),

$$mad = \frac{1}{k} \sum_{d=1}^{k} \left| \widehat{\beta}_d - \overline{\beta} \right|; \tag{4}$$

and the maximum pairwise distance among all estimates,

$$Maximum\ Distance = \left| \max_{d=1,\ldots,k} \left\{ \widehat{\beta}_d \right\} - \min_{d=1,\ldots,k} \left\{ \widehat{\beta}_d \right\} \right|. \tag{5}$$

If a particular estimator *completely* eliminated all differences across all sites, then all these distances would be exactly zero. Hence, the closer these measures get to zero, the better the performance of the estimator.

Note that, due to pure sample variation, we would never expect to see a value of zero in these measures even in settings where (2) were known to hold. To have some reference point about what can be considered reasonable values for these three measures, we present in Section 6 two sets of benchmark values based on experimental data. These benchmarks give the value of the distance measures that would be achieved by an experiment in a setting where $\beta_1 = \beta_2 = \ldots = \beta_k$ holds. The first set is derived from a "placebo" experiment, and the second exploits the availability of pre-randomization experimental data within sites. We explain both approaches in Section 6.

Finally, it is important to point out the role played by the local economic conditions (LEC) in this study. Even if we could adjust for all (observed and unobserved) *personal* characteristics of the individuals among all sites, average outcomes may fail to equalize because of differences in LEC across sites.[11] The binary treatment literature has stressed the importance of comparing individuals

[11] For instance, even if each control individual in our data had been randomly assigned to one of the seven sites, equation (2) may fail to hold because of differences in LEC across sites.

from the same local labor market when employing nonexperimental methods (e.g., Heckman et al., 1997; Heckman et al., 1998a), so the difficulty of our exercise increases as the differences in LEC across locations increase. To deal with this issue, we control for pre-randomization LEC variables by treating them as additional covariates, which is possible in our case because of the availability of different cohorts within each site.[12] We return to this issue later in the text.

3 Multiple treatment estimators

In this section we present the estimators of the parameters in (2) that we study, along with the assumptions that justify them. For reference, we consider first the *raw mean estimator*. Let $1(A)$ be the indicator function, which equals one if event A is true and zero otherwise. This estimator is then given by:

$$\widehat{\beta}_d^{raw} = \left[\sum_{i=1}^{N} Y_i 1(D_i = d)\right]\left[\sum_{i=1}^{N} 1(D_i = d)\right]^{-1}. \tag{6}$$

This estimator would be unbiased for β_d if the individuals were randomized across different locations. We use it as a measure of the initial bias, which we aim at reducing by adjusting for differences in observable characteristics across locations.

The other estimators we study are based on the following unconfoundedness or selection-on-observables assumption:

Assumption 1 (Unconfounded site)

$$1(D_i = d) \perp Y_i(0,d) \,|X_i, \text{ for all } d \in \{1,2,\ldots,k\}. \tag{7}$$

This assumption states that for all sites, and conditional on a set of covariates, the indicator variable for whether an individual belongs to a given site d is independent of her potential outcome in that site. It implies that there are no other variables related to both the indicator variable for belonging to site d and the potential outcome in that site, so that adjusting for the covariates is sufficient to remove all biases in the estimation of $\beta_d = E[Y(0,d)]$. This assumption is similar to that in Hotz et al. (2005) when comparing potential outcomes between two locations, and is referred to as *weak unconfoundedness* by Imbens (1999, 2000).[13]

As discussed by Imbens (1999), Assumption 1 is closely related to the definition of *missing at random* in the missing data literature (e.g., Rubin, 1976; Little and Rubin, 1987). For instance, suppose we are interested on learning about the population mean $E[Y(0,d)]$. The problem is that we only observe $Y_i(0,d)$ for individuals with $1(D_i = d) = 1$, while it is missing for those with $1(D_i = d) = 0$. Assumption 1 implies that conditional on X_i the two groups are comparable, so we can use the individuals with $1(D_i = d) = 1$ to learn about $E[Y(0,d)]$. Note that a key feature

[12] As we discuss in Section 6, in some specifications we also adjust the outcome for *post-randomization* LEC variables.

[13] A stronger version of Assumption 1 could be written as $D_i \perp \{Y_i(0,d)\}_{d\in\{1,2,\ldots,k\}} \,|X_i$. For a discussion of this strong version of unconfoundedness and the weak version in Assumption 1, see Imbens (1999, 2000).

of Assumption 1 is that all that matters is whether or not unit i is in site d; hence, her actual site in case she is *not* in d is not relevant.

In addition to Assumption 1, we impose an overlap assumption that guarantees that in infinite samples we are able to find individuals with the same values of the covariates across all k sites.

Assumption 2 (Simultaneous strict overlap) For all d and all x in the support of X

$$0 < \xi < \Pr\left(D_i = d | X = x\right), \text{ for some } \xi > 0. \tag{8}$$

This form of the overlap assumption is known as "strict overlap" in the binary treatment literature (e.g., Busso et al. 2009a, 2009b). The standard overlap assumption in the binary setting requires the propensity score (i.e., the probability of being in the treatment group conditional on X) to be strictly between zero and one, but otherwise allows it to be arbitrarily close to these boundaries. The strict overlap assumption plays a critical role in determining the asymptotic properties of semiparametric estimators of β_d. In particular, this condition is sufficient to guarantee $\sqrt{n}$-consistency of these estimators and finiteness of the semiparametric efficiency bound for regular estimators of β_d derived by Cattaneo (2009). In cases where the strict overlap assumption is violated, semiparametric estimators of β_d can fail to be $\sqrt{n}$-consistent (i.e., the semiparametric efficiency bound is infinite). Intuitively, as discussed by Khan and Tamer (2009) for the binary setting, if identification of β_d requires observing individuals in site d whose probability of being in site d given their covariates is arbitrarily close to zero then, although point identified, β_d cannot be estimated at the regular parametric rate ($\sqrt{n}$).[14]

The overlap condition in Assumption 2 is stronger than that of the binary treatment case, as it requires that for each individual in the population we are able to find comparable individuals in terms of covariates in each of the k sites. Intuitively, for each individual we want to learn about her potential outcomes in sites $1, \ldots, k$, but we observe only one of those k potential outcomes, so the "fundamental problem of causal inference" (Holland, 1986) is worsen. As compared to the binary treatment case, having more treatment groups implies that the conditional probabilities in (8) will be smaller in magnitude, and it also increases the probability that there is a covariate for which (8) does not hold in one of the groups. Both features represent a threat to the validity of Assumption 2, and simply reflect the difficulty of moving from a binary to a multiple treatment setting. In fact, as discussed by Imbens (1999), in the multiple treatment case there may be a particular treatment (or treatments) for which (8) fails or is close to failing. This has two effects. First, this prevents us from making precise inferences about *population* effects, since it is not possible to find comparable individuals in the rest of the treatment groups for those in the non-comparable treatment group(s). Second, it forces us to focus on the effects of those treatments for which the groups are comparable. Therefore, fewer treatment effects are estimated for a narrower population. As usual in economics,

[14] Khan and Tamer (2009) relate this class of estimators to the "identified at infinity" models (e.g., Chamberlain, 1986; Heckman, 1990), as they require some variables to take values with arbitrarily small probabilities. For more discussion on the rate of convergence of semiparametric estimators of treatment effects and its relation to the strict overlap assumption in a binary treatment setting, see Khan and Tamer (2009) and Busso et al. (2009b).

focusing on a fewer set of treatments and a narrower population reduces the external validity of the analysis, while analyzing all the treatments for the whole population despite Assumption 2 failing (or being close to failing) reduces its internal validity.

Under Assumptions 1 and 2, and using iterated expectations, we can identify β_d as:

$$\beta_d = E[E\left[Y_i|D_i = d, X_i = x\right]]. \tag{9}$$

This result suggests estimating β_d using a partial mean, which is an average of a regression function over some of its regressors while holding others fixed (Newey, 1994). In this case, the conditional expectation function of Y on d and X is estimated in a first step, and then we average this function over the covariates holding the site d fixed. The most straightforward model for the inner expectation in (9) is a linear regression of the form:

$$E\left[Y_i|D_i, X_i\right] = \sum_{j=1}^{k} \alpha_j \cdot 1\left(D_i = j\right) + \delta' X_i, \tag{10}$$

where δ is the coefficient vector for the covariates. Let the estimated coefficients in (10) be given by $\widehat{\alpha}_j$ and $\widehat{\delta}$. Then, the OLS-based estimator of β_d is given by:

$$\widehat{\beta}_d^{pmX} = \widehat{\alpha}_d + N^{-1} \textstyle\sum_{i=1}^{N} \widehat{\delta}' X_i. \tag{11}$$

In what follows, we refer to this estimator as the *partial mean linear X* estimator. We also consider a more flexible model of (10) containing polynomials of the continuous covariates and various interactions, which can be thought of as a global smoothing method (e.g., Imbens and Wooldridge, 2009). We denote this estimator by $\widehat{\beta}_d^{pmXflex}$, and refer to it as the *partial mean flexible X* estimator.

Recently, part of the focus in the program evaluation literature has been on more flexible ways to control for covariates. The main issue when controlling for the covariates without imposing any structure in the model is that, if the dimension of X is large, then nonparametric methods become intractable because of the so-called "curse of dimensionality". Rosenbaum and Rubin (1983) show that if the two potential outcomes from a binary treatment are independent of the treatment assignment conditional on X, then they are also independent conditional on the propensity score. This result implies that we only need to adjust for a scalar variable, as opposed to adjusting for all covariates.[15]

Imbens (2000) and Lechner (2001) extend the results in Rosenbaum and Rubin (1983) to the multivalued or multiple treatment setting.[16] Following Imbens (2000), define the *generalized propen-*

[15] The problem of nonparametrically estimating the regression function of the outcome on the treatment and the covariates is translated to nonparametrically estimating the propensity score. The same occurs in the multiple or multivalued treatment setting when using the generalized propensity score. For further discussion in the binary treatment setting see, for instance, Imbens (2004).

[16] One key difference between the approaches in Imbens (2000) and Lechner (2001) is that, while the latter reduces the dimension of the conditioning set from the dimension of X to the dimension of the treatment, Imbens (2000)

sity score, or GPS, as the probability of receiving a particular treatment (in our case, belonging to a particular site) conditional on the covariates:

$$r(d,x) = \Pr(D = d|X = x). \tag{12}$$

For the discussion below, it is important to keep in mind the distinction between two different random variables: the probability that an individual gets the treatment she actually received, $R_i = r_i(D_i, X_i)$, and the probability she receives a particular treatment d conditional on her covariates, $R_i^d = r_i(d, X_i)$. Clearly, $R_i^d = R_i$ for those units with $D_i = d$.

Imbens (2000) shows that under Assumptions 1 and 2 we can estimate the average potential outcomes by conditioning solely on the GPS. Analogous to the binary treatment case, he shows that $Y_i(0,d) \perp 1(D_i = d)|R_i^d$ for all $d \in \{1, 2, \ldots, k\}$. This result implies that for estimation of $E[Y(0,d)]$ is enough to compare units with $D_i \neq d$ and $D_i = d$ in terms of R_i^d.

Based on this result, Imbens (2000) proposes a partial mean approach for estimation of β_d, which can be written as:

$$\begin{aligned} &(i)\ \gamma(d,r) \equiv E[Y(0,d)|r(d,X) = r] = E[Y_i|D_i = d, R_i = r]; \\ &(ii)\ \beta_d = E[Y(0,d)] = E[\gamma(d, r_i(d, X_i))]. \end{aligned} \tag{13}$$

Therefore, the GPS can be used to estimate $\beta_d = E[Y(0,d)]$ by following the two steps in (13). First, one estimates the conditional expectation of Y as a function of D and $R = r(D,X)$. Second, to estimate β_d we average the conditional expectation $\gamma(d,r)$ over $R^d = r(d,X)$. This procedure is analogous to the partial mean approach that uses the covariates directly, see (9)-(11). However, two important differences are worth mentioning. First, contrary to that approach, we now use R_i in the regression function in the first step, and integrate over the distribution of R_i^d in the second step. Second, the inner conditional expectation $\gamma(d,r)$ does not have a causal interpretation, while the inner expectation in (9) does by Assumption 1.[17]

Hirano and Imbens (2004) implement this approach in a continuous treatment setting by estimating the regression function in the first step using a (flexible) parametric regression. Following their approach, we first estimate the regression function

$$E[Y_i|D_i, R_i] = \sum_{j=1}^{k} \alpha_j \cdot 1(D_i = j) + \sum_{j=1}^{k} [\delta_j \cdot 1(D_i = j) \cdot R_i + \eta_j \cdot 1(D_i = j) \cdot R_i^2].$$

Letting the estimated coefficients from this regression be denoted by a hat on top of the coefficient,

reduces the dimension to one, just as in the binary case.

[17] The reason $\gamma(d,r)$ does not have a casual interpretation is that, while Assumption 1 implies $1(D_i = d) \perp Y_i(0,d)|R_i^d$, for all $d \in \{1, 2, \ldots, k\}$, it does not imply $1(D_i = d) \perp Y_i(0,d)|R_i$, for all $d \in \{1, 2, \ldots, k\}$ (Imbens, 1999, 2000).

β_d is estimated as:

$$\widehat{\beta}_d^{pmGPS} = \frac{1}{N}\sum_{i=1}^{N}[\widehat{\alpha}_d \cdot 1\left(D_i = d\right) + \widehat{\delta}_d \cdot 1\left(D_i = d\right) \cdot R_i^d + \widehat{\eta}_j \cdot 1\left(D_i = d\right) \cdot (R_i^d)^2].$$

We refer to this estimator as the *GPS-based parametric partial mean estimator* of β_d.

Following Newey (1994) and more recently Flores (2007) and Flores et al. (2009), we also consider a more flexible specification in which the first-step estimator of the regression function is based on a nonparametric kernel estimator. As in Flores et al. (2009), we use a local polynomial regression of order one, which has desirable boundary properties and is commonly used in economics. Since in our case the treatment is discrete, the nonparametric regression function of Y_i on D_i and R_i in the first stage is equivalent to having one nonparametric regression function of Y_i on R_i for each site. Letting $\widehat{\gamma}\left(d, r; h\right)$ be the standard local linear estimator (e.g., Wand and Jones, 1995, p.119, eq 5.4) of $\gamma\left(d, r\right)$ in (13) based on bandwidth h, the *nonparametric partial mean estimator* of β_d is given by:

$$\widehat{\beta}_d^{pmNPR} = \frac{1}{N}\sum_{j=1}^{N}\widehat{\gamma}\left(d, R_j^d; h\right). \tag{14}$$

Note that each $\widehat{\gamma}(d, R_j^d; h)$ is obtained with R_i as the regressor using only the individuals in site d, but is evaluated for all individuals at their GPS R_j^d. In the next section we implement this estimator by using an Epanechnikov kernel, and selecting a bandwidth for each of the regressions $\gamma\left(d, r\right)$ using the procedure proposed by Fan and Gijbels (1996, p.111).[18]

In addition to employing the GPS within a partial mean framework to estimate β_d, the GPS can also be used to control for covariates using a weighting approach (e.g., Imbens, 2000; Cattaneo, 2009).[19] Similar to the binary treatment case, in a multiple treatment setting we can weight the observations receiving a given treatment level d by the probability of receiving the treatment they actually received conditional on X (i.e., R_i). More specifically, in our context we can write β_d as (Imbens, 2000):

$$\beta_d = E\left[\frac{Y_i \cdot 1\left(D_i = d\right)}{R_i}\right]. \tag{15}$$

The intuition behind weighting by R_i is creating a sample in which the covariates are balanced across all treatment arms (or sites), and then calculating the average outcome for those units with $D_i = d$ in that sample to estimate β_d. In the binary treatment literature, the weights implied by (15) are usually normalized to add to one (e.g., Imbens, 2004; Busso et al., 2009a, 2009b). Thus, the *inverse probability weighting (IPW)* estimator we use in this paper is given by

$$\widehat{\beta}_d^{ipw} = \left[\sum_{i=1}^{N}\frac{Y_i \cdot 1\left(D_i = d\right)}{R_i}\right]\left[\sum_{i=1}^{N}\frac{1\left(D_i = d\right)}{R_i}\right]^{-1}. \tag{16}$$

[18] The bandwidths are obtained by estimating the unknown terms appearing in the optimal global bandwidth using a global polynomial regression of order two. This bandwidth selection criteria has been previously used in economics; for example, in the regression discontinuity context (e.g., Lee and Lemieux, 2009).

[19] See Flores et al. (2009) for a discussion of weighting-by-the-GPS estimators in a continuous treatment setting.

Cattaneo (2009) analyzes the asymptotic properties of IPW estimators such as (16) when the GPS is nonparametrically estimated using a series-based estimator.[20,21] He shows that under certain conditions, including the simultaneous strict overlap assumption (Assumption 2), these estimators are asymptotically normal and efficient in the class of $\sqrt{n}$-consistent estimators of β_d in the sense of achieving the semiparametric efficiency bound. On the other hand, Khan and Tamer (2009) show that in many important cases IPW estimators –and more generally *any* semiparametric estimator of β_d– can fail to converge at the parametric $\sqrt{n}$ rate.[22]

Note that, similar to the binary treatment case, $\widehat{\beta}_d^{ipw}$ for $d = 1, \ldots, k$ equal the coefficients in a weighted linear regression of Y_i on the set of k dummy variables $1(D_i = j)$, with weights equal to $w_i = \sqrt{1/R_i}$. The last estimator of β_d we consider combines IPW with linear regression by adding covariates to this weighted regression.[23] It is calculated following a three-step procedure. First, we estimate the weighted regression

$$E[Y_i|D_i, X_i] = \sum_{j=1}^{k} \alpha_j \cdot 1(D_i = j) + \delta' X_i, \tag{17}$$

with weights $w_i = \sqrt{1/R_i}$. Next, we calculate the predicted value of the outcome at each site d for all individuals as $\widehat{Y}_i^d = \widehat{\alpha}_d + \widehat{\delta}' x_i$. Lastly, the estimator $\widehat{\beta}_d^{ipwX}$ is given by the weighted average of $\widehat{Y}_i^d$ using the weights w_i. We refer to this estimator in the remaining of the paper as the *IPW with covariates* estimator.[24]

In a parametric context, estimators combining IPW and linear regression share a "double robustness" property, which states that these estimators are consistent as long as either $E[Y_i|D_i, X_i]$ or the GPS is correctly specified (e.g., Robins and Rotznitzky, 1995; Scharfstein et al., 1999; Wooldridge, 2007). The first part of the argument is that if $E[Y_i|D_i, X_i]$ is correctly specified, then weighting by any nonnegative function of the covariates does not affect the consistency of (17).[25]

[20] Cattaneo (2009) uses a multinomial logistic sieve estimator for nonparametrically estimating the GPS. It generalizes the logistic sieve estimator in Hirano et al. (2003), and can be implemented in practice by simply estimating a multinomial logit with flexible functions of the covariates (e.g., polynomials and interactions).

[21] Cattaneo (2009) introduces another asymptotically efficient estimator of β_d based on the efficient influence function (EIF) for β_d, which we do not include in our study. As compared to the IPW estimator, this "EIF estimator" has the advantage of being based on the entire EIF; however, it requires the estimation of two infinite dimension parameters –the GPS and a conditional expectation which is a function of β_d. Moreover, Cattaneo (2009) does not find any meaningful differences in the results when applying both estimators to actual data, or in simulations.

[22] The rate of convergence of these semiparametric estimators depends on the tail behavior of the distribution of the covariates and the error term in the treatment equation, and in many cases can be slower than $\sqrt{n}$ (e.g., when both the regressors and the latent error term in the treatment equation are normally distributed in the binary treatment case). See Khan and Tamer (2009) and Busso et al. (2009b) for a discussion of these issues.

[23] For space considerations we do not consider other estimation methods such as blocking or matching. In a binary treatment setting, IPW estimators tend to perform better than matching and blocking estimators in simulations (e.g., Busso et al. 2009a, 2009b). In addition, in a multiple treatment setting the GPS-based estimators we consider in this paper are easier to implement and are less computationally intensive (especially the IPW estimators).

[24] Similar to the partial mean linear X estimator discussed in (11), the reason we go through this three-step procedure is to avoid using one of the sites as a reference site. In fact, the average treatment effect estimator $\widehat{\beta}_d^{ipwX} - \widehat{\beta}_f^{ipwX}$ for sites d and f is algebraically the same as the coefficient in the indicator variable for site d from the weighted regression (17) excluding the indicator variable for site f and including a constant term.

[25] In this case, however, weighting will result in a less efficient estimator relative to the unweighted regression by

Intuitively, we can think of the second part of the argument as adjusting for the covariates using linear regression in a sample in which the covariates have already been balanced by weighting by a correctly specified GPS. This would be analogous to adjusting for observed characteristics when using experimental data, which is commonly done in practice to improve precision.[26]

Finally, we also study *difference-in-difference* versions of all the estimators previously discussed. These estimators are the same as those described above, but use the outcome in differences. This specification removes any potential bias coming from temporally invariant factors correlated to both the treatment assignment (D_i) and the outcome (e.g., Heckman et al., 1997; Heckman et al., 1998a; Smith and Todd, 2005; Abadie, 2005). Hence, these estimators relax Assumption 1 by allowing the different treatment groups to also differ systematically in terms of unobserved time-invariant factors.

4 Determination of the overlap region

A key ingredient in the implementation and performance of the estimators discussed in Section 3 is the availability of comparable individuals in terms of the covariates among the different treatment groups, as implied by Assumption 2. This assumption guarantees that in infinite samples we can find comparable individuals for all units with $D_i \neq d$ in the subsample with $D_i = d$, for every site d. In finite samples, though, we may fail to find comparable units in the $D_i = d$ group. This point has been extensively analyzed in the binary treatment setting, where the propensity score plays a significant role in identifying regions of the data in which there is overlap in the covariate distributions (e.g., Heckman et al. 1997; Heckman et al., 1998a; Dehejia and Wahba, 1999, 2002; Imbens and Wooldridge, 2009). In this case, the usual approach is to focus on the "overlap" or "common support" region by dropping those individuals whose propensity score does not overlap with the propensity score of those in the other treatment arm.[27]

We propose a strategy for determining the overlap region that is motivated by a procedure commonly used in the binary treatment setting (e.g., Dehejia and Wahba, 1999, 2002). As in the binary case, we rely on the GPS R_i^d to find the set of individuals for whom there are comparable individuals in terms of covariates in each of the treatment groups. Let $R^d_{q,\{j\in A\}}$ denote the q-th quantile of the distribution of R^d over those individuals in subsample A. First, we let the overlap region with respect to a particular site (or treatment) d be given by the subsample

$$Overlap_d = \left\{ i : R_i^d \geq \max\left\{ R^d_{q,\{j:D_j=d\}}, R^d_{q,\{j:D_j\neq d\}} \right\} \right\}. \tag{18}$$

the Gauss-Markov theorem.

[26] For a discussion of this estimator in the binary treatment case see, for instance, Imbens (2004) and Imbens and Wooldridge (2009). For a formal discussion of the double robustness property in a general class of models, including estimation of average treatment effects and non-continuous outcomes (e.g., binary and count variables), see Wooldridge (2007).

[27] Unless the treatment effect is homogeneous, when implementing this type of trimming procedure one is implicitly redefining the parameter of interest to be conditional on the subpopulation with overlap in the propensity score (e.g., Crump et al., 2009; Imbens and Wooldridge, 2009).

In words, we first select the individuals whose propensity score of belonging to site d is greater than a cutoff value, which is given by the highest q-quantile from the R_i^d distribution for two groups: those individuals that are in site d and those who are not. Then, we define the *overlap* or *common support region* as the subsample given by those units that are in the overlap regions for *all* different sites *simultaneously*

$$Overlap = \bigcap_{d=1}^{k} Overlap_d. \tag{19}$$

The overlap condition for site d in (18) ensures that for every unit in the subsample $D_i \neq d$ we are able to find comparable units in the subsample $D_i = d$, which implies we can estimate $E[Y(0,d)]$ under Assumption 1.[28] Then, we take the intersection in (19) to guarantee that we estimate $E[Y(0,d)]$ $d = 1, 2, \ldots, k$ using only units that are comparable in all sites simultaneously.

This procedure to impose overlap is different from that previously used in the multiple treatment literature (e.g., Frölich et al., 2004), which uses a rule akin to (19) but defines $Overlap_d = \{i : R_i^d \in [\max_{j=1,\ldots,k}\{\min_{\{q:D_q=j\}} R_q^d\}, \min_{j=1,\ldots,k}\{\max_{\{q:D_q=j\}} R_q^d\}]\}$. The rule we propose is less stringent in two important ways. First, as implied by the weak version of the unconfoundedness assumption, it does not require the comparison of R_i^d among all k treatment groups, but only among those in groups $D_i = d$ and $D_i \neq d$. Second, it identifies individuals outside the overlap region based only on the lower tail of the distributions of R_i^d.[29]

An important issue when implementing the overlap rule (19) is selecting the quantile q that determines the amount of trimming. Even in the binary treatment literature, there is no consensus about how to select the trimming level (e.g., see Imbens and Wooldridge, 2009).[30] In this paper we set $q = 0.002$, but also present results for other quantiles ($q = 0,\ 0.001,\ 0.003,\ 0.004$ and 0.005) to evaluate the sensitivity of our results to the choice of q.

Finally, note that while the overlap rule above guarantees that for every individual in the overlap region (19) there are comparable units in each of the treatment groups, the quality of the overlap may still be poor with respect to one or more treatments. This happens, for instance, when interest lies on estimating $E[Y(0,d)]$ and, even inside the overlap region, many units with $D_i \neq d$ are comparable to only a handful of units with $D_i = d$. Hence, it is important to analyze the quality of the overlap in each of the treatment groups. In Section 6 we accomplish this by analyzing overlap plots.

[28] As discussed in Section 3, for estimation of $E[Y(0,d)]$ we only need to find comparable individuals in the subsample $D_i = d$ for those with $D_i \neq d$, and not vice versa.

[29] When estimating the average treatment effect in the binary setting, one usually looks at both tails of the distribution of the propensity score, say $\Pr(D = 1|X)$ for $D \in \{0, 1\}$. This is equivalent to analyzing the two lower tails of the distributions of $\Pr(D = 1|X)$ and $\Pr(D = 0|X)$, because $\Pr(D = 1|X) + \Pr(D = 0|X) = 1$. Since condition (18) is applied for all sites, it is not necessary to look at the upper tail of the GPS distributions.

[30] For the binary case, Crump et al. (2009) derive an efficiency-based optimal trimming rule. Characterizing an analogous rule for the multiple treatment setting is beyond the scope of this paper.

5 Data

The data used in this paper comes from the National Evaluation of Welfare-to-Work Strategies (NEWWS). This is a multi-year study conducted in the early and mid nineties to compare the effects of two approaches for helping welfare recipients (mostly single mothers) to improve their labor market outcomes and leave public assistance. The first approach emphasized labor force attachment (LFA) by encouraging participants to find employment quickly, and the second focused on human capital development (HCD) by offering academic, vocational and employment-oriented skills training. In addition, under each of these approaches the evaluation analyzed the effects of different caseload and management strategies.[31]

The programs evaluated in the NEWWS study were operated in seven sites across the U.S.: Atlanta, GA; Columbus, OH; Detroit, MI; Grand Rapids, MI; Oklahoma City; OK; Portland, OR; and Riverside, CA. In Atlanta, Grand Rapids and Riverside both LFA and HCD programs were offered, and individuals were randomly assigned to LFA, HCD or the control group. In the rest of the sites, individuals were randomized to one of the programs (LFA, HCD or a combination of both) or to the control group, which was denied access to the training services offered by the program for a pre-set "embargo" period. As explained in Section 2, we concentrate only on the individuals randomly assigned to the control group in each site in order to minimize treatment heterogeneity across sites.

We rely on the public-use version of the NEWWS data. The data contains a rich set of individual and family characteristics, information on labor market outcomes for up to five years after random assignment, and individual welfare use and labor market histories up to two years prior to random assignment. In addition, we are able to identify different cohorts in each site, although only by their *year* of random assignment in this version of the data. The year of randomization differed across sites; the earliest randomization took place in 1991, and the latest in 1994.

We use only female control individuals in the *five* sites for which all the same variables are available, and in which individuals were randomized after welfare eligibility had been determined. The final sample size in our analysis is 9,351 women: 1,372 from Atlanta; 2,037 from Detroit; 1,374 from Grand Rapids; 1,740 from Portland and 2,828 from Riverside.[32]

The outcome we use in our analysis is an indicator variable equal to one if the individual was ever employed during the two years following randomization, and zero otherwise.[33] We focus on an

[31] For a detailed description of the NEWWS study and its results see Hamilton et al. (2001).

[32] The total number of individuals in the control groups in the original seven sites is 17,521. We exclude two sites, Columbus (2,159) and Oklahoma City (4,368). Columbus has the problem of having only one year (instead of two) of labor market history prior to random assignment. We exclude it from the analysis because of the documented importance of controlling for such variables in nonexperimental settings (e.g., Heckman et al., 1997; Hotz et al., 2005). We drop Oklahoma City because in this site randomization was performed to welfare *applicants*, not to welfare *recipients* as in the remaining sites; and a large proportion (30%) of those individuals did not actually qualify for welfare. There is evidence in the literature that applicants and recipients are very different in terms of their characteristics and outcomes (e.g., Friedlander, 1988). From the individuals in the remaining five sites, we eliminate all men and individuals with unknown gender (778). In addition, we drop 431 Atlanta women for whom it is unknown whether they were embargoed from the program services during the period considered. Finally, from the remaining women, we exclude those with missing values on any of the covariates used in our analysis (434).

[33] Using a one-year employment indicator instead of the aggregate two-year indicator does not change our results

outcome measured only up to two years after random assignment because, in most sites, we cannot identify which control individuals were embargoed from receiving program services starting in year three. As mentioned in Section 3, we also consider the outcome in differences. We define it as the original outcome minus an indicator equal to one if the individual was ever employed during the two years prior to randomization. We standardize both outcomes (in levels and in differences) with respect to their respective mean and standard deviation across all sites in order to make results comparable across estimators and outcomes, and to simplify the presentation of results.[34]

The first five columns of Table 1 show the descriptive statistics of the outcomes and covariates in each site. The covariates include information on demographic and family characteristics, education, housing type and stability, welfare and food stamps use history, and earnings and employment history. As expected, there are important and usually large differences in the means of all variables across sites. For instance, while the percentage of blacks in Atlanta is 95 percent, this percentage is only 17 percent in Riverside; and while in Detroit the percentage of women with a high school or GED degree is 48 percent, it is 59 percent in Riverside and around 53 percent in the other sites In fact, all variables fail a test of equality of means across all five sites at a 5 percent significance level (not shown in the table).

As previously mentioned in Section 2, LEC play an important role in our framework, since they may prevent (2) to hold even if individual characteristics are balanced across all sites. In our data, we observe different cohorts for each site, as determined by their year of random assignment. This creates within-site variation that allows us to attempt to control for pre-randomization differences in LEC across sites. At the bottom of Table 1, we summarize the LEC faced by the individuals in different cohorts within each site, measured at the metropolitan statistical area (MSA) level. We present three variables –employment to population ratio, average real earnings and unemployment rate– that measure the LEC faced by each individual during the calendar year of random assignment, as well as their corresponding two-year growth rates in the two years prior to random assignment.[35,36] In addition, Figure 1 presents the three LEC measures for two years before and

substantively (results available upon request). We focus on an employment indicator because the public-use data available to us provides us solely with the year of random assignment and with *nominal* earnings measures. Thus, we could create only relatively rough measures of *real* earnings. In order to avoid having our results affected by how we obtain real earnings, we do not focus on earnings-related outcomes.

[34] Michalopoulos et al. (2004) use mostly the same data as us, but for binary comparisons. A key difference between their study and ours regarding the data is that they have access to a restricted-use version that allows them to compare individuals within (roughly) the same locations, which we cannot do. It also allows them to use earnings as an outcome, which we decided against (see footnote 33). Another difference is that they use Oklahoma City as a comparison group, which we do not, since in that site randomization was performed before welfare eligibility was determined (see footnote 32). A final key difference is that Michalopoulos et al. use both, "short-run" (up to two years after randomization) and "medium-run" (three to five years after randomization) outcomes. We do not use medium-run outcomes because it is not possible to guarantee, for all sites, that only individuals who were embargoed from receiving any program services over the whole five-year post-randomization period are included (see footnote 32). By including control observations that may have received services, and by keeping Oklahoma City, Michalopoulos et al. allow their control groups to potentially be "contaminated", which could seriously affect their results.

[35] Because in the public-use version of the data we only observe the *year* of random assignment, we rely on yearly LEC measures, which represent a rough approximation to the actual economic conditions faced by the program participants.

[36] The two-year growth rate of variable X between period t and $t-2$ is approximated as $\Delta \log(X) \equiv \log(X_t) -$

after randomization, by site and cohort. The differences in LEC across all sites clearly show that we are working with five distinct local labor markets. Among the five sites, the LEC in Riverside are particularly different from the rest. In Riverside, not only the LEC were the worst during the period of randomization (e.g., lowest employment to population ratio), but also the pre-randomization LEC dynamics were the worst (e.g., most negative growth rate in employment to population ratio).

6 Results

In the first subsection we discuss the estimation of the GPS and how it affects the balance of the covariates across all sites. Next, we analyze the quality of the overlap inside the common support region, and show how the extremely low overlap quality in one of the sites leads us to consider an alternative case where that site is dropped. In the third subsection we present the estimation results for those two cases, and in the last subsection we calculate some benchmarks measures based on the experimental data and perform some robustness analyses.

6.1 GPS estimation and covariate balancing

In the binary treatment setting, the propensity score is usually estimated using a logit model (e.g., Rosenbaum and Rubin, 1983; Dehejia and Wahba, 1999, 2002; Heckman et al. 1997; Heckman et al., 1998a). We follow an analogous approach and estimate the GPS using a multinomial logit model that includes all 52 individual-level covariates shown in Table 1 plus the two-year pre-randomization growth rate in employment to population ratio.[37,38]

An important property of the GPS –which does not depend on unconfoundedness– is that it balances the covariates between the individuals in a site and those not in that site, or $1\left(D_i = d\right) \perp X_i | R_i^d$ (Imbens, 1999). In the binary treatment setting this property is commonly used to gauge the adequacy of the propensity score specification (e.g., Dehejia and Wahba, 1999; Smith and Todd, 2005). We follow two strategies to examine how the estimated GPS affects the covariate balance across the different sites. Both strategies are implemented after imposing the overlap rule in (19). The first strategy tests, for each covariate, if there is *joint* equality of means across all five sites after each observation is weighted by the same weights used in the IPW estimator in (16). Summary results from these tests are presented in the first column of panel A in Table 2, which also shows equality tests for the raw means before imposing overlap. Appendix Table A1 presents the variable-by-variable results. Clearly, weighting improves the balancing among the five sites, as the number of

$\log(X_{t-2})$.

[37] It is not possible to include the LEC measures in levels into the GPS estimation because that would (almost) perfectly identify the sites. We use only one of the LEC growth measures because of high collinearity among them. Results do not change considerably when using other LEC measures.

[38] The results from this multinomial logit are not shown in the text for brevity, but are available upon request. We also considered a multinomial probit model, which has the advantage of not having the "independence from irrelevant alternatives" property. However, it has the disadvantage of being several orders of magnitude more computationally intensive. In the cases where we estimated the GPS with a multinomial probit model, it did not seem to make any sizeable difference in the results.

unbalanced covariates at the 5 percent significance level drops from 53 to 11. This reduction is not driven by an increase in variance, as weighting does bring most of the means much closer together. Not surprisingly given its large initial differences, the LEC measure is one of the variables that remains highly unbalanced, although their means are brought a little closer together by weighting.

The second strategy we use to check the covariate balance consists of a series of *pairwise* comparisons of the mean of each site versus the mean of the (pooled) remaining sites, as in Hirano and Imbens (2004). The summary results from applying this approach are presented in panel B of Table 2, while the detailed results are presented in Appendix Table A2. The results shown in Table 2 for the raw means correspond to an equality test of those two means. For a given site d, the GPS-adjusted version of these tests are obtained by dividing ("blocking") all the observations based on their value for R_i^d. The groups (or blocks) are defined by the deciles of R_i^d, which are calculated using only the units with $D_i = d$. Within each group, we calculate the difference of means in a given covariate between those individuals with $D_i = d$ and those with $D_i \neq d$. The weighted (by the number of individuals in each block) average of these differences of means is used to test the GPS-blocked equality of means between a site and all the other sites pooled. This procedure is repeated for each covariate and for all sites. Similar to our findings based on inverse probability weighting (e.g., panel A in Table 2), adjusting for the GPS improves significantly the balance of the covariates, although a few unbalanced variables remain. Overall, we conclude from Table 2 that the covariates in the raw data are highly unbalanced across sites, and that the estimated GPS when using all five sites does a reasonably good job in improving their balance.[39]

6.2 Assessment of the overlap quality

This subsection highlights the key role played by the GPS in evaluating the comparability of the different treatment groups for estimation of the parameters in (1). First, we examine the set of individuals dropped after imposing the overlap rule (19). Second, we assess the quality of overlap in our data before and after imposing the overlap rule.

A careful examination of the individuals dropped after imposing overlap can help identifying sites (or treatments groups) for which there is not sufficient overlap in the covariate distributions. In particular, it is important to look at the percentage of units dropped from each site. Having a large proportion of individuals dropped from a particular site implies that many of the units in this site are not comparable to individuals in at least one of the other sites.

In the last two rows of Table 1 we present the percentage of individuals dropped from each site after imposing the overlap condition in (19), and the overall percentage of observations dropped (almost 27 percent).[40] In Riverside, 61 percent of the observations are dropped, while in the site

[39] Following a common strategy in the binary treatment literature, we also considered other GPS specifications that included interactions and higher order terms of variables that remained unbalanced even after imposing overlap. In our case, the balancing in general did not improve under these alternative specifications, and even in some cases it worsened slightly. However, the general results presented in the paper remain virtually unchanged under those alternative specifications.

[40] In the middle five columns of Table 1 we present the descriptive statistics for the individuals *not* dropped after imposing the overlap rule. Comparing them with the descriptive statistics before imposing overlap allows us to

that follows it this number decreases considerably to 18 percent.[41] These simple statistics are already a signal that there might be insufficient overlap between Riverside and the other sites.

We now examine more closely the overlap quality in our application. A common approach for doing this in the binary treatment setting is to graph, in the same figure, the propensity score distribution of both treatment groups and look at their overlap (e.g., Dehejia and Wahba, 1999, 2002; Heckman et al. 1997; Heckman et al., 1998a). We perform an analogous analysis in Figure 2, which presents kernel density estimates of the R^d distribution for individuals with $D_i = d$ and $D_i \neq d$ for each of the five sites. As previously discussed, it is important to keep in mind that the goal when estimating β_d is to find, for every individual with $D_i \neq d$, comparable individuals in terms of R^d in the $D_i = d$ group, and not vice versa. This implies that we are interested only on the lower tail of the distributions of R^d, as stated in Assumption 2 and discussed in Section 4.

Panel A in Figure 2 presents the densities before imposing overlap. These densities show that the distributions of the corresponding GPS between the two groups of interest differ strongly for all sites, and suggest a problem of lack of overlap for some sites. This is consistent with the large differences in observable characteristics documented in Tables 1 and 2. Panel B in Figure 2 presents the densities after imposing overlap. Although it is now possible to find for every individual with $D_i \neq d$ comparable individuals in the $D_i = d$ group for all sites, the quality of the overlap remains poor in some regions of the distributions, as many observations in the first group are comparable to only few observations in the second group. As discussed below, this has important consequences on the performance of the estimators we study.

From Figure 2, it is striking how thin the overlap region in Riverside is. Before imposing overlap, the median value of the GPS R^{Riv} for those with $D_i = Riv$ is 0.95, while for those with $D_i \neq Riv$ is 0.0004 (not shown in tables). Since the conditional probabilities must add to one, this suggests that a very large fraction of individuals with $D_i = Riv$ have probabilities of being in other sites that are very small. For those units, it is very difficult to find comparable individuals in other sites. Despite dropping a very large fraction of individuals from Riverside after imposing overlap, the overlap quality in this site remains extremely poor. For instance, after imposing overlap, 75 percent of the units with $D \neq Riv$ (4,308 observations) are comparable to only 2.3 percent of the units in the $D = Riv$ group (26 observations) –not shown in tables. This implies that, even though after imposing the overlap condition (19) we guarantee that the overlap assumption is not violated in our data, we need to rely on a very small number of observations for identification of β_{Riv} over a very wide range of the support of R^{Riv}.

The extremely weak overlap for Riverside in Figure 2 reveals that, for the individuals in the other sites, there are some covariates (or combinations of them) for which it is very difficult to find comparable individuals in Riverside. Examining Appendix Tables A1 and A2, we see that two variables remain highly unbalanced even after adjusting for the GPS: the percentage of blacks in

identify the characteristics of the individuals that do not satisfy the overlap rule in each site.

[41] For reference, if instead of using $q = 0.002$ as the quantile for the overlap trimming rule, we use the weakest version of it ($q = 0$), overall we drop 12.9% of the observations; from Riverside we drop 38.4% of the observations, and for the site that follows it this number is only 3.9%.

each site and, most importantly, the LEC measure. For instance, the mean growth rates of the employment to population ratio in Portland and Riverside start about 1.79 standard deviations away from each other (see Appendix Table A1). Weighting by the GPS slightly decreases this distance to 1.36, which is still considerably high.[42] This is consistent with the large differences in LEC between Riverside and the rest of the sites documented in Section 5. Therefore, large part of the comparability issues seems to come from the LEC in Riverside being very different from those in the other sites, with the percentage of blacks playing a lesser but also important role.

The poor overlap that remains after imposing overlap in Riverside, along with the large and disproportionate percentage of observations dropped from this site, suggests performing the analyses dropping Riverside. As discussed in Section 3, the benefit of doing this is attaining a greater internal validity of the estimators, at the cost of analyzing fewer sites for a smaller population. An added benefit is that, in our context, it allows us to study the performance of the estimators when applied to locations where the LEC are relatively more similar.

For the four-site analysis, we first re-estimate the GPS using the same specification as before. The last four columns of Table 1 replicate for this case the information previously presented for the five-site case. The percentage of observations dropped when imposing the overlap rule (19) is now much lower, 12 percent, while in no site more than 22 percent of observations are dropped. Regarding balancing, the second column in Table 2 presents the summary results for the balancing tests, and Appendix Tables A3 and A4 present detailed results from these tests. As compared to the five-site case, the GPS does a better job in balancing the covariates across all groups, especially in the IPW context, where even the previously highly-unbalanced variable "black" is now balanced. Also, although still unbalanced, the growth rate of the employment to population ratio is not nearly as highly unbalanced as with five sites.

Figure 3 shows kernel densities similar to those presented for five sites. The main improvement when moving from five to four sites comes from not having a site with extremely poor overlap (i.e., Riverside). This is important since, as we show in the next subsection, it is very difficult to draw inferences about β_{Riv}. In addition, there is a small improvement in the overlap in the remaining four sites, mainly from removing observations with low values of R^d in the $D \neq d$ groups after dropping Riverside.[43] Even though in the multiple treatment setting the overlap condition is stronger, the general quality of the overlap within each site is similar to that in previous studies in a binary treatment setting (e.g., Dehejia and Wahba, 1999; Black and Smith, 2004; Smith and Todd, 2005).

[42] Similarly, Appendix Table A2 shows that the unadjusted difference in means of this variable between Riverside and the rest of the pooled sites is 1.34 standard deviations away, and it is only slightly reduced (to almost one standard deviation) after adjusting for the GPS. Not surprisingly, the t-statistic of the adjusted difference of means remains very high at 48.4 (not shown in table).

[43] For instance, note the scale change in the kernel densities for the $D \neq d$ groups.

6.3 Estimation results

In this subsection we present the results from implementing the estimators discussed in Section 3. We implement the non-GPS-based estimators before and after imposing overlap to analyze the effect of using only "comparable" units on these estimators. All the GPS-based estimators, on the other hand, were only estimated within the overlap region.[44] As previously mentioned, we use $q = 0.002$ when applying (19), and we examine the robustness of the results to alternative values of q in the next subsection.

We present first the results for the five-site case to study the performance of the estimators when the quality of the overlap is poor and the LEC differ substantially across sites. Panels A and B in Table 3 show assessment measures of the estimators for the outcome in levels and in differences, respectively. The table shows, for each estimator, the p-value from the joint equality test in (2) and the three distance measures in (3)-(5) with their corresponding 95 percent confidence interval based on 1,000 bootstrap replications. Additionally, we show the distance measures relative to those from the raw mean estimator using the outcome in levels before imposing overlap, which gives a reference level for the initial difference in outcomes across sites. In Figures 4 (levels) and 5 (differences) we show the point estimate of β_d for each site with its corresponding 95 percent confidence interval, and Appendix Table A5 presents the same information in tabular form. The dashed line in each of the graphs in Figures 4 and 5 represents the average estimate among the sites ($\overline{\beta}$), which is used to calculate the distance measures *rmsd* and *mad* –see (3) and (4). This is given as a reference point, since the goal is to equate the estimates of β_d across all five sites.

In general, the results show that the nonexperimental estimators studied are not very successful when using five sites. Only a few of them reduce the distance measures more than 50 percent with respect to the measures for the raw mean estimator in levels. Working with the outcome in differences reduces the distance measures more than using the outcome in levels, for which some of the distance measures even increase. Figures 4 and 5 make clear that the methods have difficulties in aligning β_{Riv} with the rest of the sites, as its estimate is usually well below the rest. This comes at no surprise in light of the extremely poor overlap quality in Riverside and its very different LEC.[45]

The confidence intervals for the distance measures and for many of the point estimates are

[44] For the estimators where we adjust by covariates (partial mean linear X and IPW with covariates) we include all 52 individual covariates in Table 1 plus the same LEC measure used in the GPS estimation (the two-year pre-randomization growth rate in employment to population ratio). For the partial mean flexible X estimator, we also add several higher order terms and interactions, totaling an additional 51 covariates. A complete list of these additional terms is available upon request.

[45] As Figure 1 makes clear, the differences in LEC between Riverside and the other sites exist both before and after randomization. In some specifications we adjusted the outcome by the post-randomization LEC variables. There is no clear guidance about how to do this type of adjustments. In the binary setting, Hotz et al. (2006) include post-treatment measures of LEC as additional covariates in a regression-based approach; while Galdo (2008) applies a nonparametric procedure to remove time and location fixed effects *prior* to applying matching estimators. We implemented the adjustment by running a regression of the outcome in levels against post-randomization LEC measures, and then using the residuals from this regression as an additional outcome. Probably because our yearly LEC measures are relatively rough approximations, the results from implementing our estimators to this new outcome are not significantly better than those for the outcome in differences. The results are available upon request.

very wide, in particular for the GPS-based estimators. This is a direct consequence of the overlap quality in some sites, since in some regions of the support of the GPS we are relying on a handful of observations for identification (especially in Riverside). In the binary setting, previous studies have warned about the high variance of semiparametric estimators based on the propensity score when the covariate distributions are very different between treatment and control groups (e.g., Black and Smith, 2004; Imbens and Wooldridge, 2009). When the overlap is poor, these estimators can fail to converge at the parametric $\sqrt{n}$-rate, which can lead to imprecise estimators and numerical instability (e.g., Khan and Tamer, 2009; Busso et al., 2009b). In addition, Busso et al. (2009b) document the high variance of these estimators in cases of poor overlap in a simulation study analyzing their finite sample properties in a binary treatment setting. As discussed in Section 3, in the multiple treatment setting the overlap assumption is stronger than when the treatment is binary, so this problem is likely exacerbated. On the other hand, the large standard errors of these estimators, as compared to those in more parametric models, just (correctly) reflect the uncertainty in estimating average treatment effects in cases with poor or limited overlap (e.g., Black and Smith, 2004; Imbens and Wooldridge, 2009).

We next perform the analysis for the four-site case (i.e., excluding Riverside). Table 4 presents assessment measures similar to those in Table 3, while the point estimates are presented graphically in Figures 6 and 7 and in tabular form in Appendix Table A6. The estimators now do a much better work in equalizing the estimates of β_d across the four sites, especially when using the outcome in differences. For the outcome in differences, almost all estimators reduce the three distance measures by more than 50 percent, and in many cases by about 70 percent, with respect to the raw mean estimator in levels before imposing overlap. Importantly, for most difference-in-difference estimators the joint equality test is not rejected at standard significance levels, in particular for all the GPS-based estimators and the linear regression-based partial mean with a flexible specification (after imposing overlap). Note also that in most cases imposing the overlap rule in (19) improves the performance of the linear regression-based estimators.[46]

The confidence intervals continue to be wide for the GPS-based estimators when using four sites. This is probably because, at the lower tail of the GPS distributions, we are still comparing a large number of units not in the site to a relatively small number of units in the site. Thus, the wide confidence intervals just reflect the difficulty of drawing inferences in this case. The linear regression-based estimators, on the other hand, mask this inherent uncertainty by imposing linearity assumptions that help them extrapolate to the regions of low overlap quality, resulting in narrower confidence intervals.

[46] It is important to note that the improvement in the performance of the estimators in the four-site case is driven only by the fact that Riverside is dropped, and not just by having one less site to adjust for. We repeated the four-site analysis by keeping Riverside and dropping each of the other sites one at a time. In all these four-site cases in which we kept Riverside, we encountered the same poor overlap quality in Riverside. The estimation results analogous to those in Table 4 were similar to those obtained in the five-site case in Table 3. These results are available upon request.

6.4 Experimental benchmark and robustness analysis

As discussed in Section 2, due to pure sample variation, we would never expect to see a value of zero in the distance measures in Tables 3 and 4. In order to put the results from these tables into perspective, we construct some reference levels by obtaining the value of those measures that would be achieved by an experiment, in situations where $\beta_1 = \beta_2 = \ldots = \beta_k$ is known to hold. We obtain these reference levels in two ways. First, we perform a "placebo" experiment in which we assign placebo sites randomly to the individuals in the data set, and we let $\widehat{\beta}_d$ be the sample mean in each placebo site. Table 5 presents benchmark values of the assessment measures from this exercise when applied to five and four sites using the outcome in levels and in differences.[47] Second, we use the fact that in three sites of the NEWWS study (Atlanta, Grand Rapids and Riverside) individuals were randomized to one of three possible treatments –LFA training, HCD training or a control group– within each site. Since this implies that the mean outcomes *prior* to randomization are equal for the three groups, we use a pre-randomization outcome analogous to the one actually used in our study and calculate the distance measures in (3)-(5) based on the sample mean for each group. Table 6 presents benchmark values of the assessment measures from this case.[48]

Tables 5 and 6 show benchmark values ranging from 0.020 to 0.027 for *rmsd*, 0.019 to 0.023 for *mad* and 0.048 to 0.060 for *maximum distance*. For comparison purposes, consider the benchmark values from the case with four sites in Table 5 and the IPW estimator with covariates in Table 4, both using the outcome in differences. The distance measures for the raw mean estimator in Table 4 start about four times higher than those of the benchmark case (0.082 versus 0.023). The IPW estimator substantially reduces these differences to the point that now the *rmsd* is about 30 percent higher than the benchmark of 0.023; and the *mad* and *maximum distance* are about 19 and 34 percent higher than those from the benchmark, respectively. In addition, note that if we were to take those benchmark values as fixed, and test the null hypothesis that any given distance measure from the IPW (with covariates) estimator is equal to the corresponding one from the benchmark case, we would fail to reject the hypothesis at a 5 percent significance level. We believe these results are encouraging given the difficulty of the problem at hand and the high yardstick imposed to evaluate the methods in this paper.

Finally, we examine the sensitivity of the assessment measures in Table 4 to the choice of the quantile q that determines the trimming level for the overlap rule in (19). Table 7 presents two of these measures for all the estimators, the p-value from the joint equality test in (2) and the *rmsd*, when implementing the overlap rule using the following values of q: 0, 0.001, 0.003, 0.004, and 0.005. We also include in Table 7 the results for $q = 0.002$ –which are the same as those in Table 4– to make easier the comparison of results for all the values of q considered. The percentage of

[47] The number of observations per site in each case is the same as in the original data set.

[48] The sample sizes for this exercise in Atlanta, Grand Rapids and Riverside (including the three treatments) are 4,039, 4,298 and 3,994 respectively. For Riverside, this exercise can only be conducted for women "in need of basic education", who were the only ones randomized into three treatment groups (see Hamilton et al., 2001, for details). Although the sample sizes are not the same as the ones used in Tables 3 and 4, they are of approximately the same order of magnitude; so we do not expect the distance measures to be significantly affected by these differences in sample sizes.

observations dropped in each case when imposing overlap ranges from 4.5 percent in the weakest form of the rule ($q = 0$), to 18.2 percent when $q = 0.005$. Although the point estimates vary depending on the specific choice of q –especially the GPS-based ones, which is expected given their variance–, the general points previously discussed remain valid. It is also important to note that the 95 percent confidence intervals remain remarkably stable for the different levels of q, which implies that hypothesis tests based on them are not affected noticeably by this choice.

7 Conclusion

The purpose of this paper was twofold: (i) assess the performance of unconfoundedness-based estimators of mean effects in the multiple treatment case; and (ii) analyze the role played by the GPS in evaluating the comparability of treatment groups in terms of covariates in this setting.

The overlap condition in the multiple treatment setting is stronger than in the binary case since we need to find, for each individual in the population, comparable individuals in each treatment group. This makes the estimation problem much more difficult when working with multiple treatments. Our paper highlights the crucial role played by the GPS in assessing the comparability of the different treatment groups in terms of observable characteristics. Based on the GPS, we propose a strategy to determine the overlap or common support region that is less stringent than those previously used in the multiple treatment literature, and discuss the use of the GPS when evaluating the quality of the overlap.

After we implement the proposed overlap rule to the five locations considered in this paper, we identify a site (Riverside) for which the overlap quality is extremely poor, even in the common support region. Our analysis suggests that a substantial portion of the comparability issues between Riverside and the rest of the sites comes from the large differences in local economic conditions. The case of Riverside illustrates how, in a multiple treatment setting, treatment groups for which the overlap fails or is close to failing are more likely to appear, giving rise to important trade-offs regarding the external versus internal validity of the estimators. In particular, in practice, one may choose to drop the noncomparable treatment group from the analysis and increase the internal validity of the estimators, at the cost of analyzing fewer treatments for a smaller population. This does not indicate a failure per se of the nonexperimental estimators we consider; rather, it highlights the difficulty of the problem. In our study, we assess the performance of the estimators before and after dropping Riverside from the analysis.

When we implement the unconfoundedness-based estimators using control groups with poor overlap quality and very different local economic conditions (i.e. including Riverside), they perform poorly in equalizing average outcomes across all control groups. As expected, this is mostly due to the difficulty the methods have aligning the mean outcome in Riverside with those in the rest of the sites. When the estimators are applied to control groups in locations with better overlap quality and with relatively more similar local economic conditions (i.e. dropping Riverside), their performance improves considerably. The difference-in-difference estimators perform the best and compare well to

benchmark measures derived from experimental data. The superior performance of the difference-in-difference estimators is consistent with previous findings in the binary treatment literature (e.g., Heckman et al., 1997; Heckman et al., 1998a; Smith and Todd, 2005). The improvement in the performance of the estimators when comparing individuals in more similar labor markets suggests that, when the treatment groups belong to the same local labor market, the estimators are likely to perform better.

We deem our results as encouraging regarding the performance of nonexperimental estimators for multiple treatments considering the high yardstick we impose, which entails equalizing mean outcomes across different locations. Hence, we regard the nonexperimental methods we study as valuable tools when evaluating multiple treatments in the absence of an experiment. Nevertheless, our results also show that the overlap issues that arise in the implementation of these methods can impose limits on what we can learn in any particular application. Therefore, applied researchers should always identify the common support region in their data, and pay particular attention in analyzing the overlap quality.

Finally, a natural extension of our research on unconfoundedness-based estimators of mean effects would entail using Monte Carlo simulations in order to learn more about the performance of particular estimators in different settings. More specifically, our paper points out important aspects to consider in that type of analysis, such as the degree and quality of overlap, the number of treatments considered, and external validity versus internal validity trade-offs.

References

[1] Abadie, Alberto. 2005. Semiparametric Difference-in-Differences Estimators. The Review of Economic Studies 72, no. 1 (January): 1-19.

[2] Abadie, Alberto, and Guido W. Imbens. 2006. Large Sample Properties of Matching Estimators for Average Treatment Effects. Econometrica 74, no. 1: 235-267.

[3] Behrman, Jere R., Yingmei Cheng, and Petra E. Todd. 2004. Evaluating Preschool Programs When Length of Exposure to the Program Varies: A Nonparametric Approach. The Review of Economics and Statistics 86, no. 1 (February): 108-132.

[4] Black, Dan A., and Jeffrey A. Smith. How robust is the evidence on the effects of college quality? Evidence from matching. Journal of Econometrics 121, no. 1-2 (July-August): 99-124.

[5] Busso, Matias, John DiNardo, and Justin McCrary. 2009a. New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators. IZA Discussion Paper no. 3998 (February).

[6] Busso, Matias, John DiNardo, and Justin McCrary. 2009b. Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects. University of Michigan, Department of Economics (June).

[7] Cattaneo, Matias. 2009. Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability. University of Michigan, Department of Economics (July).

[8] Chamberlain, Gary. 1986. Asymptotic efficiency in semi-parametric models with censoring. Journal of Econometrics 32, no. 2 (July): 189-218.

[9] Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. Biometrika 96, no. 1 (March): 187-199.

[10] Dehejia, Rajeev. 2005. Practical propensity score matching: a reply to Smith and Todd. Journal of Econometrics 125, no. 1-2 (March-April): 355-364.

[11] Dehejia, Rajeev H., and Sadek Wahba. 1999. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. Journal of the American Statistical Association 94, no. 448 (December): 1053-1062.

[12] Dehejia, Rajeev H., and Sadek Wahba. 2002. Propensity Score-Matching Methods for Nonexperimental Causal Studies. The Review of Economics and Statistics 84, no. 1 (February): 151-161.

[13] Fan, Jianqing, and Irène Gijbels. 1996. Local polynomial modelling and its applications. New York: Chapman & Hall/CRC Press.

[14] Flores, Carlos A. 2007. Estimation of Dose-Response Functions and Optimal Doses with a Continuous Treatment. University of Miami, Department of Economics (November).

[15] Flores, Carlos A., Alfonso Flores-Lagunes, Arturo Gonzalez, and Todd C. Neumann. 2009. Estimating the Effects of Length of Exposure to a Training Program: The Case of Job Corps. University of Miami, Department of Economics (March).

[16] Fraker, Thomas, and Rebecca Maynard. 1987. The Adequacy of Comparison Group Designs for Evaluation of Employment-Related Programs. Journal of Human Resources 22, no. 2 (Spring): 194-227.

[17] Friedlander, Daniel. 1988. Subgroup Impacts and Performance Indicators for Selected Welfare Employment Programs. New York: Manpower Demonstration Research Corporation.

[18] Friedlander, Daniel, and Philip K. Robins. 1995. Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods. The American Economic Review 85, no. 4 (September): 923-937.

[19] Frölich, Markus. 2004a. Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators. The Review of Economics and Statistics 86, no. 1 (February): 77-90.

[20] Frölich, Markus. 2004b. Programme Evaluation with Multiple Treatments. Journal of Economic Surveys 18, no. 2 (April): 181-224.

[21] Frölich, Markus, Almas Heshmati, and Michael Lechner. 2004. A microeconometric evaluation of rehabilitation of long-term sickness in Sweden. Journal of Applied Econometrics 19, no. 3 (May/June): 375-396.

[22] Galdo, Jose. 2008. Treatment Effects for Profiling Unemployment Insurance Programs: Semiparametric Estimation of Matching Models with Fixed Effects. Carleton University, School of Public Policy and Administration (December).

[23] Hahn, Jinyong. 1998. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. Econometrica 66, no. 2 (March): 315-331.

[24] Hamilton, Gayle, Stephen Freedman, Lisa Gennetian, Charles Michalopoulos, Johana Walter, Diana Adams-Ciardullo, Ann Gassman-Pines, Sharon McGroder, Martha Zaslow, Jennifer

Brooks, and Surjeet Ahluwalia. 2001. How Effective are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs. Washington, D.C.: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation and Administration for Children and Families, and U.S. Department of Education.

[25] Heckman, James. 1990. Varieties of Selection Bias. The American Economic Review 80, no. 2 (May): 313-318.

[26] Heckman, James J., and V. Joseph Hotz. 1989. Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training (with discussion). Journal of the American Statistical Association 84, no. 408 (December): 862-874.

[27] Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998a. Characterizing Selection Bias Using Experimental Data. Econometrica 66, no. 5 (September): 1017-1098.

[28] Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. The Review of Economic Studies 64, no. 4: 605-654.

[29] Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1998b. Matching As An Econometric Evaluation Estimator. The Review of Economic Studies 65, no. 2 (April): 261-294.

[30] Heckman, James J., Robert J. Lalonde, and Jeffrey A. Smith. 1999. The economics and econometrics of active labor market programs. In Handbook of Labor Economics, ed. Orley C. Ashenfelter and David Card, Volume 3, Part 1:1865-2097. Elsevier.

[31] Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. Econometrica 71, no. 4 (July): 1161-1189.

[32] Hirano, Keisuke, and Guido W. Imbens. 2004. The Propensity Score with Continuous Treatments. In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, ed. Andrew Gelman and Xiao-Li Meng, 73-84. Hoboken, NJ: John Wiley and Sons.

[33] Holland, Paul W. 1986. Statistics and Causal Inference (with discussion). Journal of the American Statistical Association 81, no. 396 (December): 945-970.

[34] Hotz, V. Joseph, Guido W. Imbens, and Jacob A. Klerman. 2006. Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program. Journal of Labor Economics 24, no. 3 (July): 521-566.

[35] Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. Predicting the efficacy of future training programs using past experiences at other locations. Journal of Econometrics 125, no. 1-2 (March-April): 241-270.

[36] Imai, Kosuke, and David A. van Dyk. 2004. Causal Inference With General Treatment Regimes. Journal of the American Statistical Association 99, no. 467: 854-866.

[37] Imbens, Guido W. 1999. The Role of the Propensity Score in Estimating Dose-Response Functions. NBER Technical Working Paper Series no. 237 (April).

[38] Imbens, Guido W. 2000. The Role of the Propensity Score in Estimating Dose-Response Functions. Biometrika 87, no. 3 (September): 706-710.

[39] Imbens, Guido W. 2004. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. The Review of Economics and Statistics 86, no. 1: 4-29.

[40] Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. Recent Developments in the Econometrics of Program Evaluation. Journal of Economic Literature 47, no. 1 (March): 5-86.

[41] Khan, Shakeeb, and Elie Tamer. Irregular Identification, Support Conditions and Inverse Weight Estimation. Northwestern University, Department of Economics (July).

[42] Kluve, Jochen, Hilmar Schneider, Arne Uhlendorff, and Zhong Zhao. 2007. Evaluating Continuous Training Programs Using the Generalized Propensity Score. IZA Discussion Paper, no. 3255 (December).

[43] LaLonde, Robert J. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. The American Economic Review 76, no. 4 (September): 604-620.

[44] Lechner, Michael. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In Econometric Evaluation of Labour Market Policies, ed. Michael Lechner and Friedhelm Pfeiffer, 43-58. ZEW Economic Studies 13. New York: Springer-Verlag.

[45] Lechner, Michael. 2002a. Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods. Journal of the Royal Statistical Society, Series A (Statistics in Society) 165, no. 1: 59-82.

[46] Lechner, Michael. 2002b. Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies. Review of Economics and Statistics 84, no. 2 (May): 205-220.

[47] Lee, David S., and Thomas Lemieux. 2009. Regression Discontinuity Designs in Economics. NBER Working Paper Series no. 14723 (February).

[48] Little, Roderick J. A., and Donald B. Rubin. 1987. Statistical analysis with missing data. New York: John Wiley and Sons.

[49] Michalopoulos, Charles, Howard S. Bloom, and Carolyn J. Hill. 2004. Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs? The Review of Economics and Statistics 86, no. 1 (February): 156-179.

[50] Mitnik, Oscar A. 2008. Intergenerational transmission of welfare dependency: The effects of length of exposure. University of Miami, Department of Economics (March).

[51] Mueser, Peter R., Kenneth R. Troske, and Alexey Gorislavsky. 2007. Using State Administrative Data to Measure Program Performance. The Review of Economics and Statistics 89, no. 4 (November): 761-783.

[52] Newey, Whitney K. 1994. Kernel Estimation of Partial Means and a General Variance Estimator. Econometric Theory 10, no. 2 (June): 233-253.

[53] Neyman, Jerzy. 1923. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9 translated by D. M. Dabrowska and T. P. Speed. 1990. Statistical Science 5, no. 4 (November): 465-472.

[54] Plesca, Miana, and Jeffrey A. Smith. 2007. Evaluating multi-treatment programs: theory and evidence from the U.S. Job Training Partnership Act experiment. Empirical Economics 32, no. 2 (May): 491-528.

[55] Robins, James M., and Andrea Rotnitzky. 1995. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. Journal of the American Statistical Association 90, no. 429 (March): 122-129.

[56] Rosenbaum, Paul R., and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, no. 1 (April): 41-55.

[57] Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, no. 5 (October): 688-701.

[58] Rubin, Donald B. 1976. Inference and missing data. Biometrika 63, no. 3 (December): 581-592.

[59] Scharfstein, Daniel O., Andrea Rotnitzky, and James M. Robins. 1999. Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. Journal of the American Statistical Association 94, no. 448 (December): 1096-1120.

[60] Smith, Jeffrey A., and Petra E. Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics 125, no. 1-2 (March-April): 305-353.

[61] Wand, M. P., and M. C. Jones. 1995. Kernel smoothing. New York: Chapman & Hall/CRC Press.

[62] Wooldridge, Jeffrey M. 2007. Inverse probability weighted estimation for general missing data problems. Journal of Econometrics 141, no. 2 (December): 1281-1301.

[63] Zhao, Zhong. 2004. Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence. The Review of Economics and Statistics 86, no. 1 (February): 91-107.

Table 1. Descriptive statistics for the NEWWS data

Variables	Before imposing overlap					After imposing overlap - 5 sites					After imposing overlap - 4 sites			
	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR
Outcomes														
Ever employed in 2 years after RA	0.59	0.59	0.70	0.59	0.46	0.58	0.59	0.68	0.58	0.42	0.58	0.59	0.68	0.58
	(0.49)	(0.49)	(0.46)	(0.49)	(0.50)	(0.49)	(0.49)	(0.47)	(0.49)	(0.49)	(0.49)	(0.49)	(0.47)	(0.49)
Ever employed in 2 years after RA (in diff)	0.07	0.15	0.03	0.05	-0.07	0.08	0.15	0.03	0.06	-0.05	0.08	0.15	0.03	0.06
	(0.57)	(0.57)	(0.55)	(0.57)	(0.58)	(0.57)	(0.57)	(0.56)	(0.58)	(0.57)	(0.57)	(0.57)	(0.56)	(0.59)
Covariates														
Demographic and family characteristics														
Black	0.95	0.89	0.41	0.20	0.17	0.94	0.90	0.47	0.24	0.23	0.95	0.90	0.47	0.25
	(0.22)	(0.32)	(0.49)	(0.40)	(0.38)	(0.23)	(0.30)	(0.50)	(0.43)	(0.42)	(0.23)	(0.30)	(0.50)	(0.43)
Age 30-39 years old	0.51	0.35	0.29	0.40	0.45	0.50	0.35	0.31	0.42	0.47	0.50	0.35	0.32	0.43
	(0.50)	(0.48)	(0.46)	(0.49)	(0.50)	(0.50)	(0.48)	(0.46)	(0.49)	(0.50)	(0.50)	(0.48)	(0.47)	(0.50)
Age 40+ years old	0.14	0.11	0.09	0.08	0.13	0.12	0.11	0.09	0.09	0.13	0.13	0.11	0.09	0.09
	(0.34)	(0.32)	(0.28)	(0.27)	(0.34)	(0.33)	(0.31)	(0.29)	(0.29)	(0.34)	(0.33)	(0.32)	(0.29)	(0.29)
Became mother as a teenager	0.45	0.45	0.51	0.34	0.35	0.45	0.45	0.51	0.34	0.36	0.45	0.46	0.51	0.34
	(0.50)	(0.50)	(0.50)	(0.47)	(0.48)	(0.50)	(0.50)	(0.50)	(0.47)	(0.48)	(0.50)	(0.50)	(0.50)	(0.47)
Never married	0.62	0.69	0.58	0.49	0.34	0.64	0.70	0.59	0.51	0.39	0.63	0.70	0.58	0.50
	(0.48)	(0.46)	(0.49)	(0.50)	(0.47)	(0.48)	(0.46)	(0.49)	(0.50)	(0.49)	(0.48)	(0.46)	(0.49)	(0.50)
Any child 0-5 years old	0.42	0.65	0.69	0.71	0.58	0.45	0.66	0.67	0.68	0.60	0.45	0.65	0.66	0.67
	(0.49)	(0.48)	(0.46)	(0.46)	(0.49)	(0.50)	(0.47)	(0.47)	(0.47)	(0.49)	(0.50)	(0.48)	(0.47)	(0.47)
Any child 6-12 years old	0.70	0.48	0.43	0.52	0.59	0.71	0.49	0.46	0.56	0.61	0.70	0.49	0.46	0.57
	(0.46)	(0.50)	(0.49)	(0.50)	(0.49)	(0.45)	(0.50)	(0.50)	(0.50)	(0.49)	(0.46)	(0.50)	(0.50)	(0.50)
Two children in household	0.34	0.30	0.36	0.33	0.32	0.34	0.30	0.37	0.33	0.33	0.34	0.30	0.36	0.34
	(0.47)	(0.46)	(0.48)	(0.47)	(0.47)	(0.47)	(0.46)	(0.48)	(0.47)	(0.47)	(0.47)	(0.46)	(0.48)	(0.47)
Three or more children in household	0.31	0.27	0.19	0.30	0.28	0.33	0.28	0.21	0.31	0.31	0.33	0.27	0.21	0.31
	(0.46)	(0.44)	(0.39)	(0.46)	(0.45)	(0.47)	(0.45)	(0.40)	(0.46)	(0.46)	(0.47)	(0.45)	(0.40)	(0.46)
Education characteristics														
10th grade	0.14	0.15	0.13	0.17	0.11	0.14	0.14	0.13	0.17	0.13	0.14	0.14	0.13	0.17
	(0.35)	(0.35)	(0.34)	(0.38)	(0.31)	(0.35)	(0.35)	(0.34)	(0.38)	(0.34)	(0.35)	(0.35)	(0.34)	(0.37)
11th grade	0.17	0.25	0.20	0.22	0.18	0.19	0.26	0.21	0.20	0.17	0.18	0.26	0.20	0.20
	(0.38)	(0.44)	(0.40)	(0.41)	(0.38)	(0.39)	(0.44)	(0.41)	(0.40)	(0.38)	(0.39)	(0.44)	(0.40)	(0.40)
Grade 12 or higher	0.57	0.50	0.54	0.45	0.57	0.56	0.50	0.53	0.45	0.51	0.56	0.50	0.54	0.46
	(0.49)	(0.50)	(0.50)	(0.50)	(0.49)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Highest degree = High School or GED	0.53	0.48	0.54	0.53	0.59	0.53	0.49	0.53	0.53	0.52	0.52	0.48	0.54	0.53
	(0.50)	(0.50)	(0.50)	(0.50)	(0.49)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Housing type and housing stability														
Lives in public/subsidized house	0.59	0.07	0.16	0.29	0.09	0.62	0.07	0.18	0.33	0.18	0.60	0.07	0.18	0.34
	(0.49)	(0.26)	(0.37)	(0.46)	(0.29)	(0.49)	(0.26)	(0.39)	(0.47)	(0.38)	(0.49)	(0.26)	(0.38)	(0.47)
One or two moves in past 2 years	0.49	0.48	0.51	0.47	0.54	0.49	0.48	0.53	0.47	0.50	0.49	0.48	0.53	0.47
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
3 or more moves in past 2 years	0.08	0.08	0.25	0.23	0.22	0.08	0.07	0.21	0.19	0.18	0.08	0.07	0.21	0.18
	(0.27)	(0.27)	(0.43)	(0.42)	(0.41)	(0.27)	(0.26)	(0.41)	(0.39)	(0.38)	(0.28)	(0.26)	(0.41)	(0.39)
Welfare use history														
On welfare for less than 2 years	0.26	0.23	0.38	0.32	0.44	0.24	0.23	0.34	0.29	0.33	0.24	0.23	0.35	0.29
	(0.44)	(0.42)	(0.49)	(0.47)	(0.50)	(0.43)	(0.42)	(0.48)	(0.46)	(0.47)	(0.43)	(0.42)	(0.48)	(0.46)

(continues in next page)

Table 1. Descriptive statistics for the NEWWS data (continuation)

Variables	Before imposing overlap					After imposing overlap - 5 sites					After imposing overlap - 4 sites			
	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR
On welfare for 2-5 years	0.25	0.25	0.31	0.35	0.28	0.26	0.26	0.32	0.36	0.33	0.26	0.25	0.32	0.36
	(0.43)	(0.44)	(0.46)	(0.48)	(0.45)	(0.44)	(0.44)	(0.47)	(0.48)	(0.47)	(0.44)	(0.44)	(0.47)	(0.48)
On welfare 5-10 years	0.24	0.24	0.17	0.23	0.16	0.26	0.24	0.18	0.24	0.19	0.25	0.24	0.18	0.24
	(0.43)	(0.43)	(0.38)	(0.42)	(0.37)	(0.44)	(0.43)	(0.39)	(0.43)	(0.39)	(0.43)	(0.43)	(0.39)	(0.43)
Received welfare in Q1 before RA	0.97	0.90	0.77	0.79	0.73	0.96	0.91	0.85	0.88	0.88	0.96	0.91	0.84	0.88
	(0.18)	(0.29)	(0.42)	(0.41)	(0.44)	(0.19)	(0.28)	(0.35)	(0.33)	(0.33)	(0.18)	(0.28)	(0.36)	(0.32)
Received welfare in Q2 before RA	0.93	0.86	0.70	0.74	0.49	0.92	0.86	0.78	0.83	0.76	0.92	0.86	0.75	0.83
	(0.26)	(0.35)	(0.46)	(0.44)	(0.50)	(0.27)	(0.34)	(0.42)	(0.38)	(0.42)	(0.27)	(0.34)	(0.43)	(0.38)
Received welfare in Q3 before RA	0.85	0.84	0.68	0.72	0.46	0.84	0.85	0.75	0.78	0.72	0.84	0.85	0.73	0.78
	(0.36)	(0.37)	(0.47)	(0.45)	(0.50)	(0.36)	(0.36)	(0.44)	(0.41)	(0.45)	(0.36)	(0.36)	(0.45)	(0.41)
Received welfare in Q4 before RA	0.73	0.83	0.67	0.69	0.44	0.77	0.83	0.72	0.74	0.67	0.77	0.83	0.70	0.73
	(0.44)	(0.38)	(0.47)	(0.46)	(0.50)	(0.42)	(0.37)	(0.45)	(0.44)	(0.47)	(0.42)	(0.37)	(0.46)	(0.44)
Received welfare in Q5 before RA	0.69	0.81	0.64	0.64	0.41	0.73	0.81	0.69	0.68	0.62	0.73	0.81	0.67	0.68
	(0.46)	(0.40)	(0.48)	(0.48)	(0.49)	(0.44)	(0.39)	(0.46)	(0.47)	(0.49)	(0.45)	(0.39)	(0.47)	(0.47)
Received welfare in Q6 before RA	0.66	0.79	0.61	0.61	0.39	0.69	0.79	0.65	0.65	0.57	0.69	0.79	0.64	0.64
	(0.47)	(0.41)	(0.49)	(0.49)	(0.49)	(0.46)	(0.41)	(0.48)	(0.48)	(0.49)	(0.46)	(0.40)	(0.48)	(0.48)
Received welfare in Q7 before RA	0.64	0.77	0.56	0.58	0.37	0.68	0.77	0.60	0.61	0.55	0.67	0.77	0.59	0.61
	(0.48)	(0.42)	(0.50)	(0.49)	(0.48)	(0.47)	(0.42)	(0.49)	(0.49)	(0.50)	(0.47)	(0.42)	(0.49)	(0.49)
Food stamps use history														
Received FS in Q1 before RA	0.97	0.94	0.85	0.86	0.62	0.98	0.95	0.90	0.91	0.83	0.97	0.95	0.89	0.92
	(0.17)	(0.23)	(0.36)	(0.35)	(0.48)	(0.15)	(0.22)	(0.30)	(0.28)	(0.37)	(0.17)	(0.22)	(0.31)	(0.27)
Received FS in Q2 before RA	0.95	0.89	0.76	0.81	0.42	0.95	0.90	0.83	0.87	0.75	0.95	0.90	0.81	0.87
	(0.22)	(0.31)	(0.43)	(0.39)	(0.49)	(0.21)	(0.31)	(0.37)	(0.33)	(0.43)	(0.23)	(0.31)	(0.39)	(0.33)
Received FS in Q3 before RA	0.90	0.87	0.72	0.79	0.39	0.91	0.88	0.79	0.84	0.71	0.90	0.87	0.76	0.83
	(0.30)	(0.34)	(0.45)	(0.41)	(0.49)	(0.28)	(0.33)	(0.41)	(0.37)	(0.46)	(0.30)	(0.33)	(0.43)	(0.37)
Received FS in Q4 before RA	0.83	0.86	0.72	0.76	0.36	0.86	0.87	0.77	0.80	0.65	0.85	0.86	0.74	0.80
	(0.38)	(0.35)	(0.45)	(0.43)	(0.48)	(0.35)	(0.34)	(0.42)	(0.40)	(0.48)	(0.36)	(0.34)	(0.44)	(0.40)
Received FS in Q5 before RA	0.78	0.83	0.67	0.72	0.33	0.83	0.84	0.72	0.75	0.61	0.81	0.83	0.70	0.75
	(0.42)	(0.38)	(0.47)	(0.45)	(0.47)	(0.38)	(0.37)	(0.45)	(0.43)	(0.49)	(0.39)	(0.37)	(0.46)	(0.43)
Received FS in Q6 before RA	0.75	0.81	0.64	0.70	0.31	0.80	0.81	0.68	0.73	0.56	0.78	0.81	0.67	0.73
	(0.43)	(0.40)	(0.48)	(0.46)	(0.46)	(0.40)	(0.39)	(0.46)	(0.45)	(0.50)	(0.41)	(0.39)	(0.47)	(0.45)
Received FS in Q7 before RA	0.72	0.78	0.60	0.66	0.29	0.78	0.79	0.63	0.68	0.53	0.75	0.79	0.63	0.69
	(0.45)	(0.41)	(0.49)	(0.47)	(0.45)	(0.42)	(0.41)	(0.48)	(0.46)	(0.50)	(0.43)	(0.41)	(0.48)	(0.46)
Employment history														
Employed in Q1 before RA	0.18	0.18	0.29	0.23	0.22	0.18	0.18	0.27	0.20	0.17	0.18	0.18	0.27	0.19
	(0.39)	(0.38)	(0.45)	(0.42)	(0.41)	(0.38)	(0.38)	(0.44)	(0.40)	(0.38)	(0.39)	(0.38)	(0.44)	(0.40)
Employed in Q2 before RA	0.18	0.18	0.30	0.25	0.25	0.18	0.18	0.27	0.22	0.18	0.18	0.18	0.27	0.22
	(0.38)	(0.38)	(0.46)	(0.43)	(0.43)	(0.38)	(0.38)	(0.44)	(0.41)	(0.39)	(0.38)	(0.38)	(0.44)	(0.41)
Employed in Q3 before RA	0.19	0.18	0.29	0.25	0.26	0.19	0.17	0.27	0.22	0.20	0.19	0.17	0.27	0.22
	(0.39)	(0.38)	(0.46)	(0.43)	(0.44)	(0.39)	(0.38)	(0.44)	(0.42)	(0.40)	(0.39)	(0.38)	(0.44)	(0.42)
Employed in Q4 before RA	0.22	0.17	0.30	0.24	0.28	0.21	0.17	0.27	0.22	0.21	0.20	0.17	0.27	0.22
	(0.41)	(0.38)	(0.46)	(0.42)	(0.45)	(0.40)	(0.38)	(0.45)	(0.41)	(0.41)	(0.40)	(0.38)	(0.45)	(0.41)
Employed in Q5 before RA	0.24	0.17	0.31	0.24	0.28	0.23	0.17	0.29	0.23	0.22	0.23	0.17	0.29	0.23
	(0.43)	(0.38)	(0.46)	(0.43)	(0.45)	(0.42)	(0.38)	(0.45)	(0.42)	(0.42)	(0.42)	(0.38)	(0.45)	(0.42)

(continues in next page)

Table 1. Descriptive statistics for the NEWWS data (continuation)

Variables	Before imposing overlap					After imposing overlap - 5 sites					After imposing overlap - 4 sites			
	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR
Employed in Q6 before RA	0.27	0.18	0.34	0.25	0.29	0.25	0.18	0.32	0.23	0.23	0.25	0.18	0.32	0.24
	(0.44)	(0.38)	(0.47)	(0.43)	(0.45)	(0.43)	(0.38)	(0.47)	(0.42)	(0.42)	(0.43)	(0.39)	(0.47)	(0.43)
Employed in Q7 before RA	0.29	0.18	0.36	0.26	0.29	0.27	0.18	0.35	0.25	0.22	0.27	0.18	0.35	0.25
	(0.45)	(0.39)	(0.48)	(0.44)	(0.45)	(0.44)	(0.39)	(0.48)	(0.43)	(0.42)	(0.45)	(0.39)	(0.48)	(0.43)
Employed in Q8 before RA	0.30	0.18	0.39	0.27	0.30	0.28	0.18	0.37	0.26	0.25	0.28	0.18	0.37	0.27
	(0.46)	(0.38)	(0.49)	(0.45)	(0.46)	(0.45)	(0.38)	(0.48)	(0.44)	(0.44)	(0.45)	(0.38)	(0.48)	(0.44)
Employed at RA (self reported)	0.07	0.07	0.13	0.09	0.13	0.07	0.07	0.13	0.08	0.10	0.07	0.07	0.13	0.08
	(0.26)	(0.25)	(0.34)	(0.28)	(0.33)	(0.26)	(0.25)	(0.33)	(0.27)	(0.30)	(0.26)	(0.25)	(0.34)	(0.27)
Ever worked FT 6+ months at same job	0.72	0.46	0.64	0.77	0.71	0.72	0.47	0.63	0.76	0.69	0.72	0.47	0.64	0.76
	(0.45)	(0.50)	(0.48)	(0.42)	(0.45)	(0.45)	(0.50)	(0.48)	(0.43)	(0.46)	(0.45)	(0.50)	(0.48)	(0.43)
Earnings history (real $ /1,000)														
Earnings Q1 before RA	0.23	0.21	0.36	0.33	0.43	0.23	0.21	0.32	0.26	0.29	0.23	0.21	0.32	0.25
	(0.82)	(0.68)	(1.06)	(0.89)	(1.23)	(0.76)	(0.69)	(1.03)	(0.74)	(0.98)	(0.75)	(0.69)	(1.02)	(0.74)
Earnings Q2 before RA	0.26	0.25	0.52	0.41	0.63	0.26	0.25	0.45	0.32	0.36	0.26	0.25	0.44	0.32
	(0.85)	(0.82)	(1.29)	(1.04)	(1.55)	(0.82)	(0.82)	(1.23)	(0.90)	(1.06)	(0.81)	(0.82)	(1.22)	(0.91)
Earnings Q3 before RA	0.29	0.26	0.55	0.41	0.72	0.28	0.24	0.46	0.32	0.37	0.28	0.25	0.47	0.32
	(0.92)	(0.89)	(1.33)	(1.07)	(1.73)	(0.86)	(0.79)	(1.25)	(0.93)	(1.09)	(0.85)	(0.81)	(1.24)	(0.94)
Earnings Q4 before RA	0.41	0.25	0.53	0.43	0.74	0.39	0.24	0.47	0.37	0.43	0.38	0.25	0.48	0.37
	(1.22)	(0.82)	(1.29)	(1.14)	(1.68)	(1.18)	(0.82)	(1.26)	(1.06)	(1.20)	(1.16)	(0.82)	(1.26)	(1.06)
Earnings Q5 before RA	0.51	0.29	0.57	0.46	0.79	0.45	0.28	0.53	0.41	0.51	0.45	0.29	0.54	0.41
	(1.27)	(0.94)	(1.32)	(1.16)	(1.82)	(1.15)	(0.92)	(1.31)	(1.08)	(1.36)	(1.14)	(0.93)	(1.31)	(1.09)
Earnings Q6 before RA	0.62	0.31	0.62	0.51	0.80	0.55	0.31	0.58	0.48	0.52	0.56	0.32	0.59	0.49
	(1.44)	(1.01)	(1.41)	(1.26)	(1.81)	(1.31)	(1.01)	(1.42)	(1.24)	(1.36)	(1.35)	(1.02)	(1.42)	(1.26)
Earnings Q7 before RA	0.72	0.32	0.68	0.54	0.83	0.61	0.32	0.65	0.52	0.56	0.63	0.33	0.66	0.53
	(1.65)	(1.06)	(1.44)	(1.31)	(1.89)	(1.47)	(1.06)	(1.44)	(1.31)	(1.40)	(1.50)	(1.06)	(1.45)	(1.33)
Earnings Q8 before RA	0.74	0.33	0.69	0.57	0.85	0.65	0.33	0.67	0.55	0.61	0.67	0.34	0.68	0.56
	(1.61)	(1.09)	(1.45)	(1.35)	(1.86)	(1.48)	(1.11)	(1.48)	(1.34)	(1.53)	(1.52)	(1.11)	(1.47)	(1.37)
Any earnings year before RA (self reported)	0.23	0.20	0.46	0.36	0.40	0.23	0.20	0.42	0.31	0.30	0.23	0.20	0.42	0.31
	(0.42)	(0.40)	(0.50)	(0.48)	(0.49)	(0.42)	(0.40)	(0.49)	(0.46)	(0.46)	(0.42)	(0.40)	(0.49)	(0.46)
Local economic conditions														
Employment/population year of RA	0.52	0.46	0.49	0.49	0.29	0.52	0.46	0.49	0.49	0.29	0.52	0.45	0.49	0.49
Average total earnings year of RA ($1000)	32.39	35.90	29.12	30.00	27.80	32.37	35.88	29.12	29.98	27.68	32.37	35.88	29.12	29.98
Unemployment rate year of RA	5.93	7.38	7.42	5.36	10.45	5.89	7.40	7.45	5.46	10.53	5.91	7.41	7.49	5.48
Emp/pop growth rate 2 yrs before RA (Δ logs)	-0.03	0.00	-0.02	0.00	-0.05	-0.03	0.00	-0.02	0.00	-0.05	-0.03	0.00	-0.02	0.00
Avg erns grwth rate 2 yrs before RA (Δ logs)	0.03	0.03	0.01	0.02	-0.01	0.03	0.03	0.01	0.02	-0.01	0.03	0.03	0.01	0.02
Unemp growth rate 2 yrs before RA (Δ logs)	0.22	-0.24	0.11	-0.06	0.42	0.22	-0.24	0.12	-0.03	0.36	0.22	-0.24	0.13	-0.02
Number of observations per site	1,372	2,037	1,374	1,740	2,828	1,184	1,943	1,185	1,432	1,107	1,245	1,945	1,193	1,360
Total number of observations			9,351					6,851					5,743	
Obs dropped per site due to overlap (%)						13.7%	4.6%	13.8%	17.7%	60.9%	9.3%	4.5%	13.2%	21.8%
Total obs dropped due to overlap (%)								26.7%					12.0%	

Table 2. Summary results from covariate-balancing analysis

A. Joint equality of means tests for each covariate across all sites

Method	Number of covariates for which p-value ≤ 0.05	
	5 sites	4 sites
Raw means before overlap	53	52
GPS-based Inverse Probability Weighting	11	5
Total number of covariates	53	53

B. Difference of means tests for each covariate - Each site versus all other sites pooled

Method	Number of covariates for which p-value ≤ 0.05	
	5 sites	4 sites
Raw means before overlap		
Atlanta vs others	43	36
Detroit vs others	50	47
Grand Rapids vs others	35	49
Portland vs others	37	34
Riverside vs others	49	-
Blocking on GPS		
Atlanta vs others	4	1
Detroit vs others	6	2
Grand Rapids vs others	1	1
Portland vs others	4	6
Riverside vs others	2	-
Total number of covariates	53	53

Note: GPS-based balancing tests are applied only to observations that satisfy the overlap condition.

Table 3. Assesment measures of estimators of the average employment rate in two years after random assignment - 5 sites

Estimator	P-value joint equality Wald test	Distance measures					
		Root Mean Square Distance		Mean Absolute Distance		Maximum Distance	
		Estimated value	Relative to Raw Mean	Estimated value	Relative to Raw Mean	Estimated value	Relative to Raw Mean
A. Outcome in levels							
Raw Mean - No Ovlp	0.000	0.153	1.000	0.101	1.000	0.481	1.000
		[0.137,0.170]		[0.095,0.121]		[0.427,0.534]	
Raw Mean - Ovlp	0.000	0.171	1.122	0.123	1.210	0.530	1.101
		[0.150,0.194]		[0.108,0.143]		[0.459,0.599]	
Linear regression-based							
Partial Mean Linear X - No Ovlp	0.000	0.111	0.726	0.083	0.819	0.329	0.684
		[0.093,0.133]		[0.070,0.106]		[0.269,0.393]	
Partial Mean Linear X - Ovlp	0.000	0.107	0.698	0.081	0.797	0.308	0.639
		[0.089,0.137]		[0.068,0.109]		[0.250,0.392]	
Partial Mean Flex X - No Ovlp	0.000	0.113	0.742	0.086	0.850	0.329	0.684
		[0.095,0.136]		[0.073,0.109]		[0.269,0.390]	
Partial Mean Flex X - Ovlp	0.000	0.108	0.706	0.083	0.818	0.304	0.632
		[0.089,0.137]		[0.069,0.110]		[0.253,0.388]	
GPS-based (imposing Ovlp)							
Parametric Partial Mean	0.016	0.117	0.769	0.093	0.920	0.332	0.689
		[0.076,0.193]		[0.061,0.153]		[0.209,0.547]	
Nonparametric Partial Mean	0.016	0.150	0.984	0.119	1.177	0.396	0.822
		[0.094,0.234]		[0.075,0.184]		[0.261,0.650]	
IPW No Covariates	0.005	0.165	1.083	0.129	1.271	0.448	0.930
		[0.112,0.258]		[0.092,0.229]		[0.304,0.705]	
IPW With Covariates	0.009	0.138	0.906	0.110	1.082	0.371	0.770
		[0.097,0.226]		[0.080,0.181]		[0.275,0.619]	
B. Outcome in differences (with respecto to years 1 and 2 before RA)							
Raw Estimator - No Ovlp	0.000	0.121	0.795	0.091	0.898	0.376	0.780
		[0.108,0.139]		[0.079,0.110]		[0.329,0.427]	
Raw Estimator - Ovlp	0.000	0.116	0.759	0.090	0.890	0.355	0.738
		[0.103,0.143]		[0.077,0.113]		[0.308,0.434]	
Linear regression-based							
Partial Mean Linear X - No Ovlp	0.000	0.074	0.485	0.056	0.554	0.217	0.451
		[0.054,0.101]		[0.043,0.082]		[0.154,0.290]	
Partial Mean Linear X - Ovlp	0.001	0.071	0.463	0.055	0.545	0.216	0.449
		[0.053,0.106]		[0.042,0.088]		[0.152,0.311]	
Partial Mean Flex X - No Ovlp	0.000	0.100	0.652	0.079	0.776	0.285	0.591
		[0.083,0.121]		[0.065,0.098]		[0.232,0.342]	
Partial Mean Flex X - Ovlp	0.000	0.092	0.602	0.073	0.725	0.253	0.525
		[0.076,0.120]		[0.060,0.099]		[0.213,0.336]	
GPS-based (imposing Ovlp)							
Parametric Partial Mean	0.027	0.107	0.698	0.078	0.766	0.318	0.661
		[0.068,0.173]		[0.056,0.135]		[0.189,0.500]	
Nonparametric Partial Mean	0.017	0.122	0.801	0.087	0.855	0.379	0.788
		[0.087,0.196]		[0.069,0.152]		[0.240,0.593]	
IPW No Covariates	0.043	0.116	0.763	0.090	0.885	0.321	0.666
		[0.075,0.205]		[0.063,0.162]		[0.210,0.604]	
IPW With Covariates	0.763	0.062	0.406	0.044	0.434	0.192	0.398
		[0.040,0.161]		[0.034,0.129]		[0.105,0.472]	

Note: Bootstrap confidence intervals in brackets (based on 1,000 replications).

Table 4. Assesment measures of estimators of the average employment rate in two years after random assignment - 4 sites

Estimator	P-value joint equality Wald test	Distance measures: Root Mean Square Distance Estimated value	Root Mean Square Distance Relative to Raw Mean	Mean Absolute Distance Estimated value	Mean Absolute Distance Relative to Raw Mean	Maximum Distance Estimated value	Maximum Distance Relative to Raw Mean
A. Outcome in levels							
Raw Mean - No Ovlp	0.000	0.097	1.000	0.084	1.000	0.228	1.000
		[0.078,0.119]		[0.066,0.101]		[0.194,0.296]	
Raw Mean - Ovlp	0.000	0.084	0.868	0.072	0.862	0.206	0.902
		[0.064,0.109]		[0.054,0.092]		[0.162,0.282]	
Linear regression-based							
Partial Mean Linear X - No Ovlp	0.012	0.043	0.443	0.037	0.437	0.114	0.500
		[0.028,0.069]		[0.023,0.061]		[0.071,0.181]	
Partial Mean Linear X - Ovlp	0.112	0.036	0.369	0.033	0.388	0.093	0.409
		[0.021,0.067]		[0.018,0.060]		[0.054,0.173]	
Partial Mean Flex X - No Ovlp	0.011	0.046	0.470	0.043	0.511	0.116	0.510
		[0.028,0.073]		[0.025,0.067]		[0.073,0.184]	
Partial Mean Flex X - Ovlp	0.098	0.039	0.398	0.037	0.446	0.093	0.408
		[0.024,0.072]		[0.020,0.065]		[0.062,0.181]	
GPS-based (imposing Ovlp)							
Parametric Partial Mean	0.037	0.060	0.617	0.047	0.561	0.166	0.728
		[0.034,0.100]		[0.030,0.086]		[0.088,0.262]	
Nonparametric Partial Mean	0.637	0.031	0.317	0.026	0.312	0.077	0.337
		[0.021,0.087]		[0.018,0.076]		[0.053,0.229]	
IPW No Covariates	0.621	0.056	0.583	0.048	0.575	0.151	0.661
		[0.029,0.168]		[0.025,0.141]		[0.073,0.433]	
IPW With Covariates	0.284	0.058	0.597	0.045	0.532	0.162	0.712
		[0.028,0.113]		[0.025,0.096]		[0.069,0.291]	
B. Outcome in differences (with respecto to years 1 and 2 before RA)							
Raw Estimator - No Ovlp	0.000	0.082	0.851	0.067	0.803	0.217	0.953
		[0.066,0.103]		[0.055,0.087]		[0.171,0.276]	
Raw Estimator - Ovlp	0.000	0.082	0.850	0.066	0.790	0.222	0.973
		[0.065,0.105]		[0.053,0.088]		[0.169,0.284]	
Linear regression-based							
Partial Mean Linear X - No Ovlp	0.104	0.042	0.430	0.034	0.411	0.103	0.451
		[0.021,0.074]		[0.018,0.063]		[0.054,0.190]	
Partial Mean Linear X - Ovlp	0.040	0.050	0.515	0.042	0.503	0.128	0.562
		[0.030,0.086]		[0.025,0.075]		[0.076,0.221]	
Partial Mean Flex X - No Ovlp	0.050	0.034	0.351	0.028	0.333	0.093	0.409
		[0.020,0.058]		[0.017,0.050]		[0.052,0.152]	
Partial Mean Flex X - Ovlp	0.249	0.028	0.294	0.023	0.278	0.073	0.319
		[0.018,0.060]		[0.015,0.052]		[0.045,0.155]	
GPS-based (imposing Ovlp)							
Parametric Partial Mean	0.629	0.031	0.322	0.026	0.314	0.080	0.350
		[0.018,0.083]		[0.015,0.070]		[0.045,0.216]	
Nonparametric Partial Mean	0.462	0.041	0.428	0.036	0.425	0.103	0.453
		[0.022,0.100]		[0.019,0.085]		[0.058,0.259]	
IPW No Covariates	0.788	0.030	0.305	0.025	0.302	0.072	0.315
		[0.025,0.130]		[0.021,0.110]		[0.065,0.335]	
IPW With Covariates	0.894	0.030	0.306	0.025	0.299	0.076	0.335
		[0.017,0.091]		[0.015,0.081]		[0.042,0.239]	

Note: Bootstrap confidence intervals in brackets (based on 1,000 replications).

Table 5. Benchmark values of the assessment measures for Raw Mean estimator from placebo experiments
Outcome: Employment rate in two years after random assignment

Outcome	P-value joint equality Wald test	Distance measures: Root Mean Square Distance	Distance measures: Mean Absolute Distance	Distance measures: Maximum Distance
A. 5 sites				
Levels	0.436	0.020 [0.013,0.043]	0.020 [0.011,0.037]	0.048 [0.035,0.122]
DID	0.158	0.027 [0.018,0.051]	0.024 [0.015,0.045]	0.064 [0.046,0.141]
B. 4 sites				
Levels	0.491	0.020 [0.010,0.047]	0.019 [0.009,0.042]	0.048 [0.027,0.120]
DID	0.344	0.023 [0.013,0.048]	0.021 [0.010,0.043]	0.056 [0.032,0.126]

Note: Bootstrap confidence intervals in brackets (based on 1,000 replications).

Table 6. Benchmark values of the assessment measures for Raw Mean estimator from using within-site experimental treatment groups (3 treatments per site)
Outcome: Employment rate in two years *prior* to random assignment

Site	P-value joint equality Wald test	Distance measures: Root Mean Square Distance	Distance measures: Mean Absolute Distance	Distance measures: Maximum Distance
ATL	0.270	0.024 [0.009,0.051]	0.023 [0.008,0.046]	0.052 [0.022,0.120]
GRP	0.250	0.024 [0.009,0.051]	0.021 [0.008,0.045]	0.057 [0.021,0.120]
RIV	0.283	0.025 [0.009,0.055]	0.023 [0.008,0.049]	0.060 [0.021,0.129]

Note: Bootstrap confidence intervals in brackets (based on 1,000 replications).

Table 7. Assesment measures of estimators when applying different overlap trimming rules (quantile q) - 4 sites
Outcome: Employment Rate in Two Years after Random Assignment

	Overlap rule: q=0.000		Overlap rule: q=0.001		Overlap rule: q=0.002		Overlap rule: q=0.003		Overlap rule: q=0.004		Overlap rule: q=0.005	
Estimator	**P-value jnt equality Wald test**	**Root Mean Square Distance**	**P-value jnt equality Wald test**	**Root Mean Square Distance**	**P-value jnt equality Wald test**	**Root Mean Square Distance**	**P-value jnt equality Wald test**	**Root Mean Square Distance**	**P-value jnt equality Wald test**	**Root Mean Square Distance**	**P-value jnt equality Wald test**	**Root Mean Square Distance**
A. Outcome in levels												
Raw Mean - No Ovlp	0.000	0.097 [0.078,0.119]	0.000	0.097 [0.078,0.119]	0.000	0.097 [0.078,0.119]	0.000	0.097 [0.078,0.119]	0.000	0.097 [0.078,0.119]	0.000	0.097 [0.078,0.119]
Raw Mean - Ovlp	0.000	0.093 [0.070,0.114]	0.000	0.087 [0.066,0.112]	0.000	0.084 [0.064,0.109]	0.000	0.081 [0.062,0.109]	0.000	0.082 [0.060,0.107]	0.000	0.081 [0.059,0.107]
Linear regression-based												
Partial Mean Linear X - No Ovlp	0.012	0.043 [0.028,0.069]	0.012	0.043 [0.028,0.069]	0.012	0.043 [0.028,0.069]	0.012	0.043 [0.028,0.069]	0.012	0.043 [0.028,0.069]	0.012	0.043 [0.028,0.069]
Partial Mean Linear X - Ovlp	0.058	0.039 [0.023,0.068]	0.088	0.037 [0.023,0.067]	0.112	0.036 [0.021,0.067]	0.094	0.039 [0.020,0.066]	0.115	0.038 [0.019,0.065]	0.106	0.039 [0.019,0.065]
Partial Mean Flex X - No Ovlp	0.011	0.046 [0.028,0.073]	0.011	0.046 [0.028,0.073]	0.011	0.046 [0.028,0.073]	0.011	0.046 [0.028,0.073]	0.011	0.046 [0.028,0.073]	0.011	0.046 [0.028,0.073]
Partial Mean Flex X - Ovlp	0.055	0.041 [0.025,0.072]	0.077	0.040 [0.025,0.073]	0.098	0.039 [0.024,0.072]	0.061	0.044 [0.023,0.072]	0.069	0.043 [0.022,0.072]	0.056	0.045 [0.023,0.072]
GPS-based (imposing Ovlp)												
Parametric Partial Mean	0.016	0.067 [0.041,0.106]	0.062	0.057 [0.037,0.102]	0.037	0.060 [0.034,0.100]	0.070	0.055 [0.034,0.098]	0.079	0.055 [0.034,0.098]	0.084	0.054 [0.033,0.097]
Nonparametric Partial Mean	0.467	0.038 [0.020,0.090]	0.687	0.029 [0.020,0.088]	0.637	0.031 [0.021,0.087]	0.685	0.029 [0.021,0.090]	0.691	0.029 [0.022,0.089]	0.650	0.031 [0.022,0.093]
IPW No Covariates	0.360	0.089 [0.036,0.191]	0.725	0.043 [0.030,0.186]	0.621	0.056 [0.029,0.168]	0.626	0.049 [0.026,0.146]	0.490	0.064 [0.025,0.143]	0.626	0.049 [0.027,0.139]
IPW With Covariates	0.200	0.066 [0.031,0.125]	0.463	0.043 [0.028,0.118]	0.284	0.058 [0.028,0.113]	0.351	0.050 [0.027,0.111]	0.187	0.066 [0.026,0.110]	0.280	0.057 [0.024,0.108]
B. Outcome in differences (with respecto to years 1 and 2 before RA												
Raw Mean - No Ovlp	0.000	0.082 [0.066,0.103]	0.000	0.082 [0.066,0.103]	0.000	0.082 [0.066,0.103]	0.000	0.082 [0.066,0.103]	0.000	0.082 [0.066,0.103]	0.000	0.082 [0.066,0.103]
Raw Mean - Ovlp	0.000	0.082 [0.065,0.104]	0.000	0.081 [0.065,0.105]	0.000	0.082 [0.065,0.105]	0.000	0.085 [0.065,0.105]	0.000	0.082 [0.064,0.106]	0.000	0.081 [0.064,0.105]
Linear regression-based												
Partial Mean Linear X - No Ovlp	0.104	0.042 [0.021,0.074]	0.104	0.042 [0.021,0.074]	0.104	0.042 [0.021,0.074]	0.104	0.042 [0.021,0.074]	0.104	0.042 [0.021,0.074]	0.104	0.042 [0.021,0.074]
Partial Mean Linear X - Ovlp	0.046	0.049 [0.028,0.084]	0.036	0.051 [0.029,0.084]	0.040	0.050 [0.030,0.086]	0.017	0.056 [0.031,0.087]	0.018	0.056 [0.031,0.089]	0.014	0.058 [0.032,0.090]
Partial Mean Flex X - No Ovlp	0.050	0.034 [0.020,0.058]	0.050	0.034 [0.020,0.058]	0.050	0.034 [0.020,0.058]	0.050	0.034 [0.020,0.058]	0.050	0.034 [0.020,0.058]	0.050	0.034 [0.020,0.058]
Partial Mean Flex X - Ovlp	0.131	0.032 [0.018,0.059]	0.160	0.032 [0.017,0.059]	0.249	0.028 [0.018,0.060]	0.141	0.034 [0.016,0.059]	0.141	0.034 [0.016,0.059]	0.127	0.036 [0.015,0.059]
GPS-based (imposing Ovlp)												
Parametric Partial Mean	0.738	0.026 [0.017,0.078]	0.635	0.030 [0.017,0.079]	0.629	0.031 [0.018,0.083]	0.656	0.030 [0.019,0.085]	0.582	0.034 [0.019,0.087]	0.590	0.034 [0.020,0.088]
Nonparametric Partial Mean	0.319	0.047 [0.024,0.099]	0.449	0.041 [0.023,0.098]	0.462	0.041 [0.022,0.100]	0.475	0.041 [0.025,0.099]	0.442	0.043 [0.024,0.101]	0.466	0.043 [0.024,0.102]
IPW No Covariates	0.729	0.038 [0.024,0.122]	0.754	0.035 [0.026,0.126]	0.788	0.030 [0.025,0.130]	0.812	0.027 [0.026,0.121]	0.668	0.045 [0.027,0.120]	0.667	0.043 [0.026,0.119]
IPW With Covariates	0.991	0.012 [0.016,0.086]	0.990	0.010 [0.017,0.091]	0.894	0.030 [0.017,0.091]	0.955	0.019 [0.020,0.095]	0.734	0.042 [0.020,0.099]	0.708	0.041 [0.020,0.099]
Sample size after overlap	6,228		5,888		5,743		5,545		5,393		5,337	
Obs dropped due to overlap (%)	4.5%		9.7%		12.0%		15.0%		17.3%		18.2%	

Notes: Results based on 1,000 bootstrap replications.

Figure 1. Local economic conditions

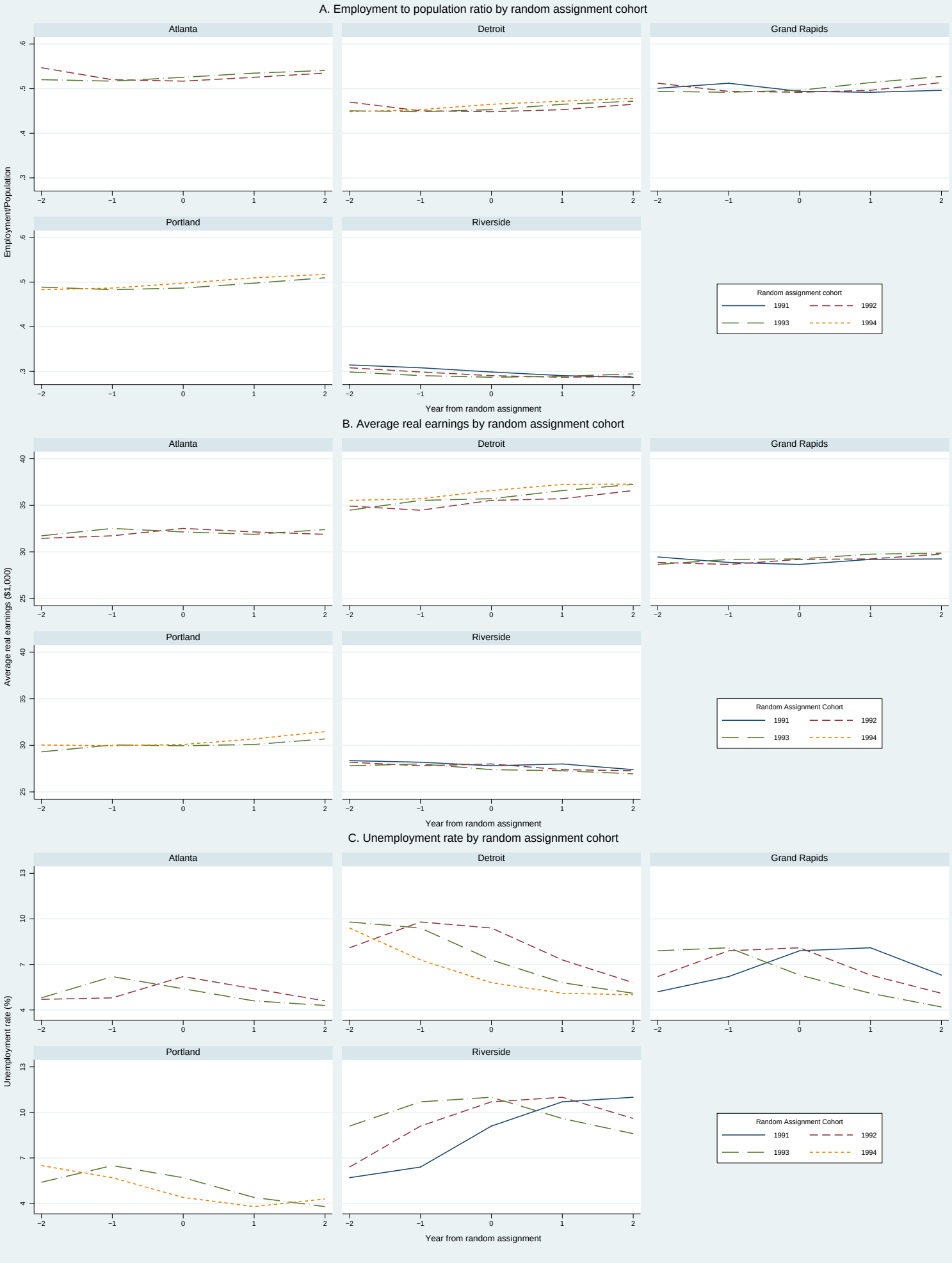
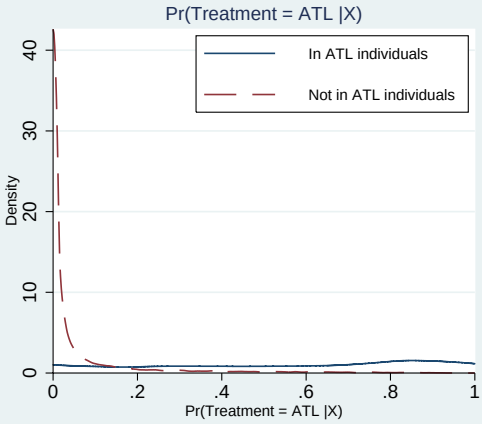
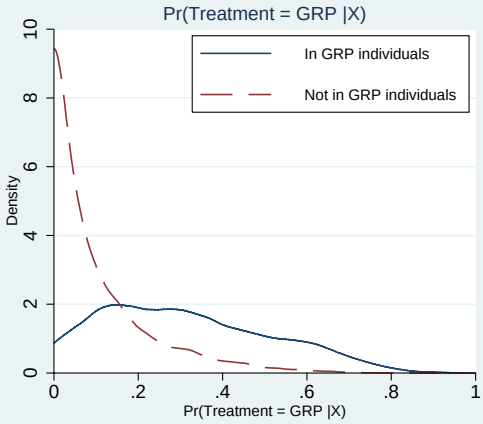
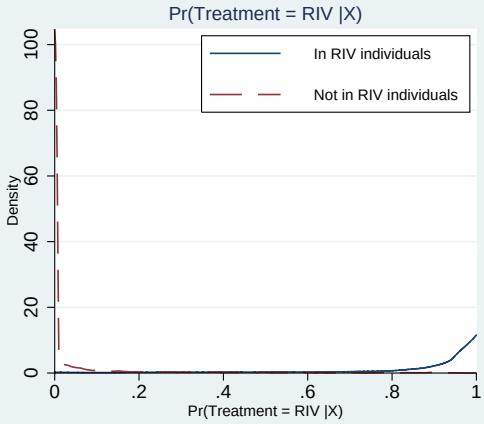


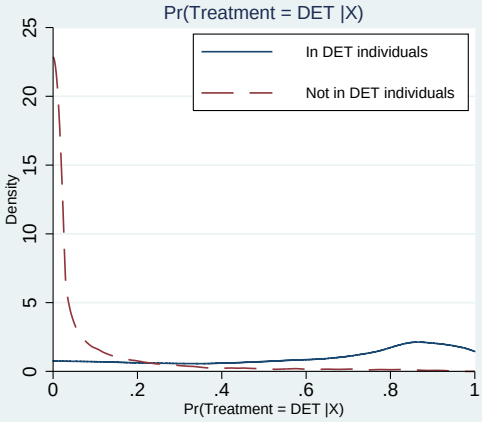
Figure 2. Kernel densities of estimated GPS − 5 sites

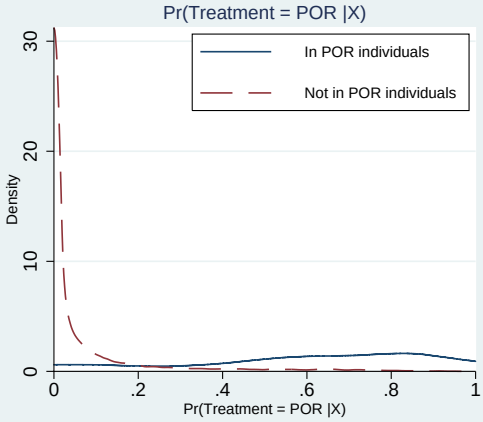
A. Before imposing overlap



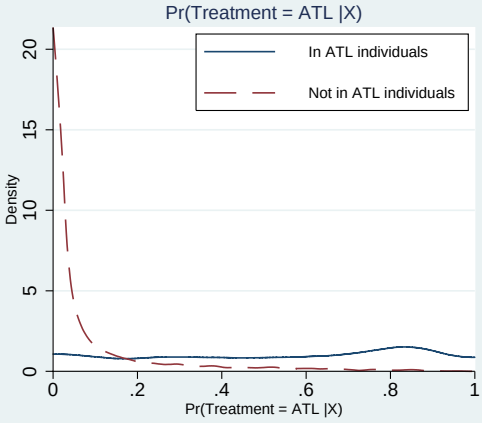


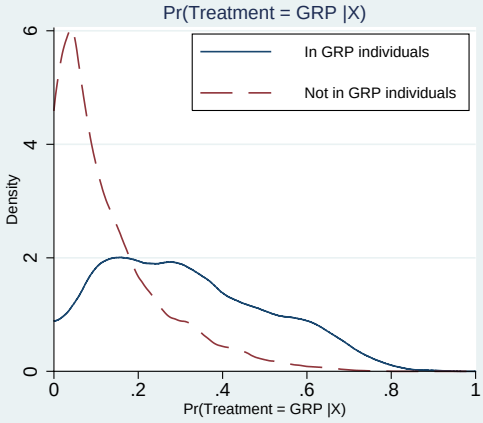


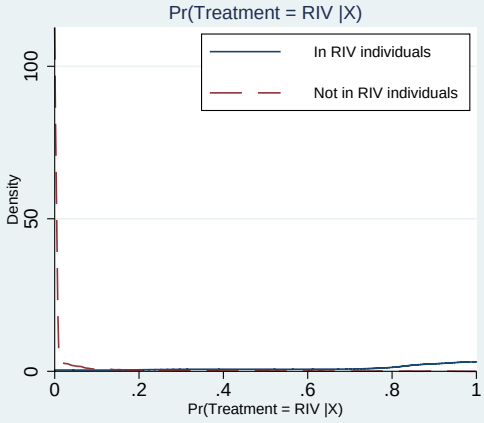


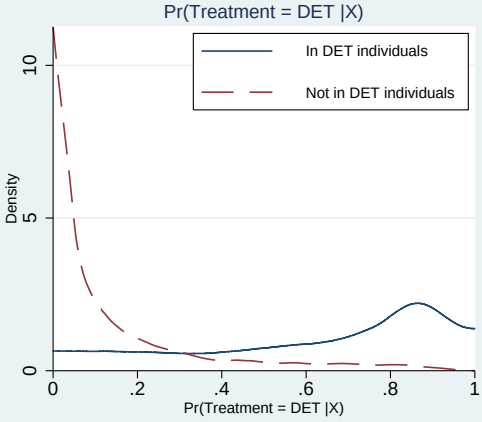


B. After imposing overlap









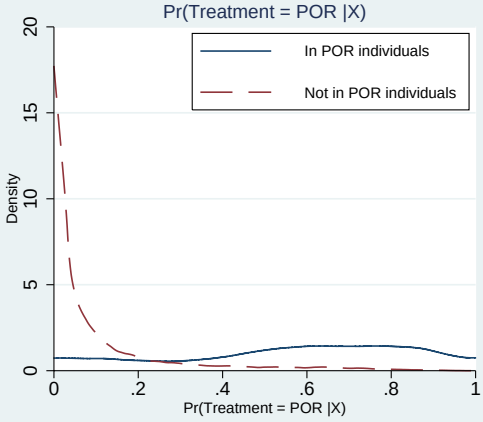


Figure 3. Kernel densities of estimated GPS − 4 sites

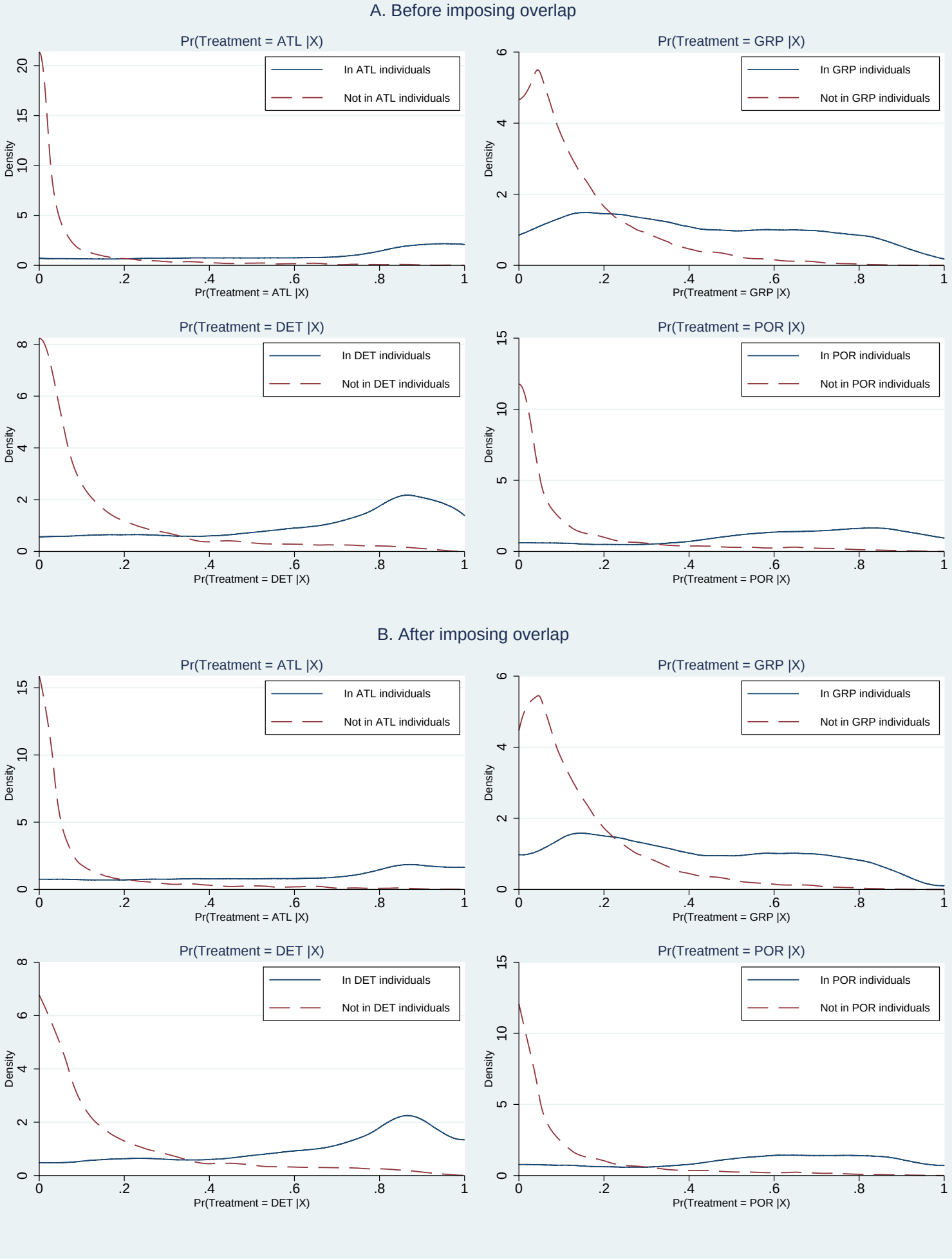


Figure 4. Estimated mean outcome in levels − 5 Sites

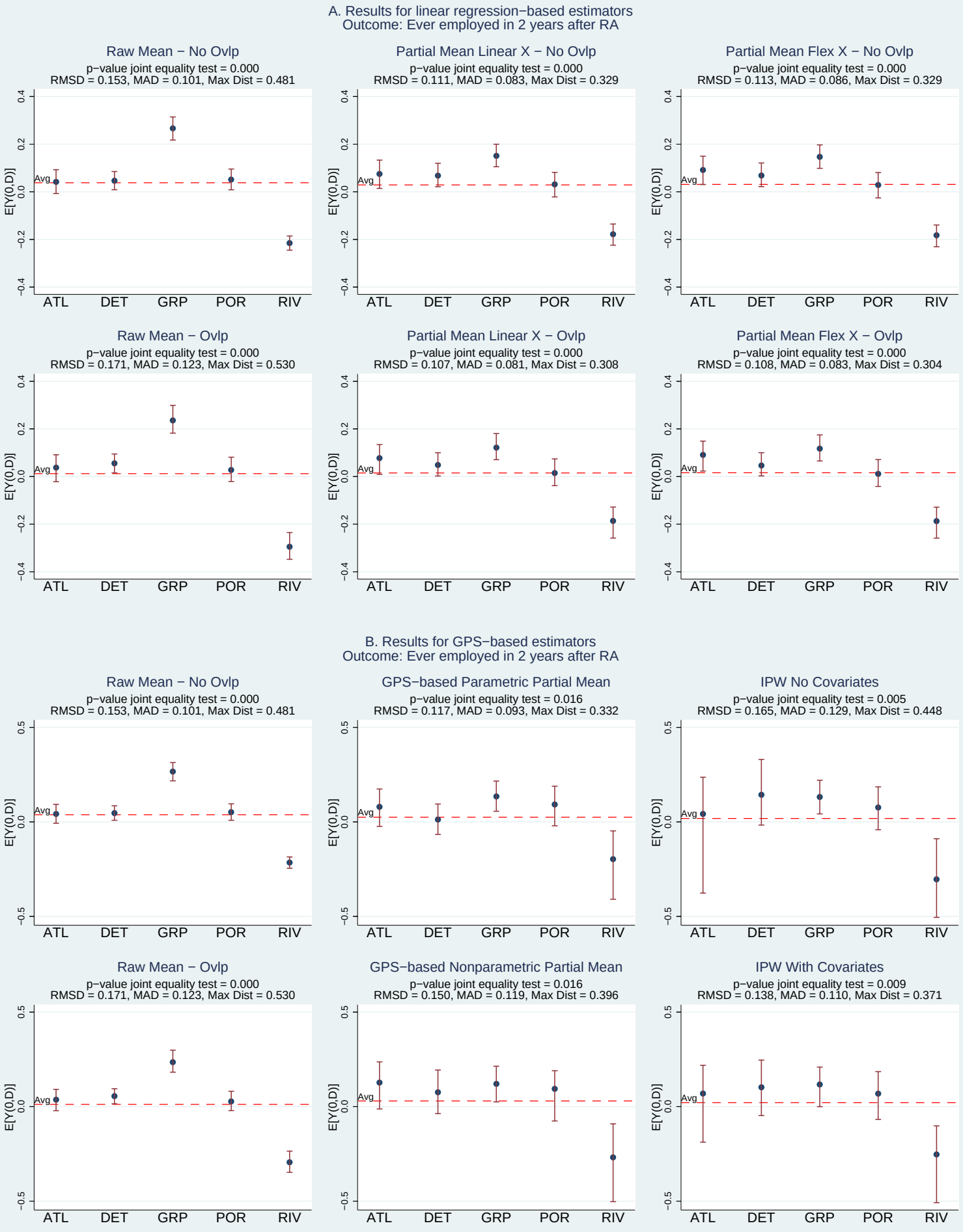


Figure 5. Estimated mean outcome in differences − 5 Sites

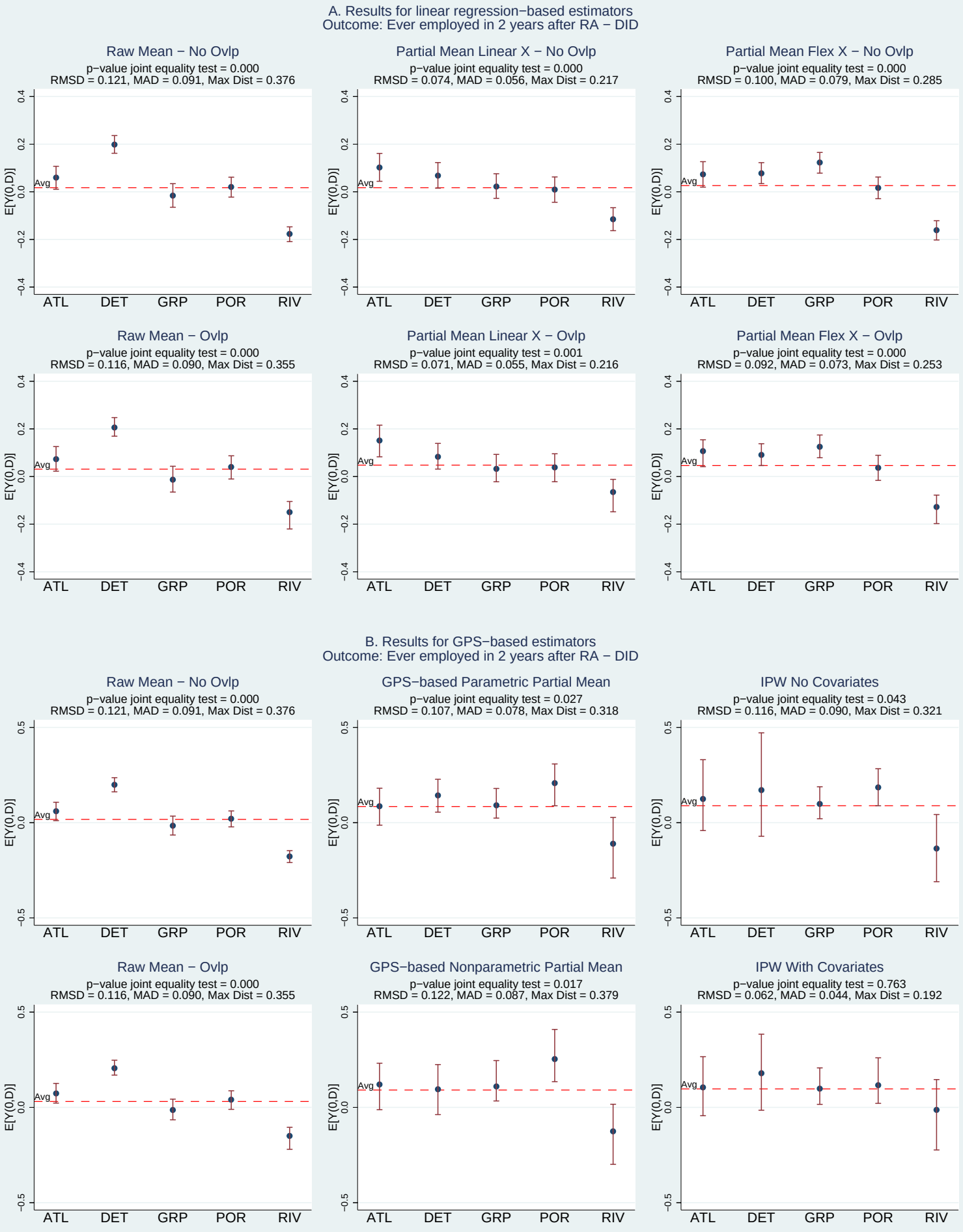


Figure 6. Estimated mean outcome in levels − 4 Sites

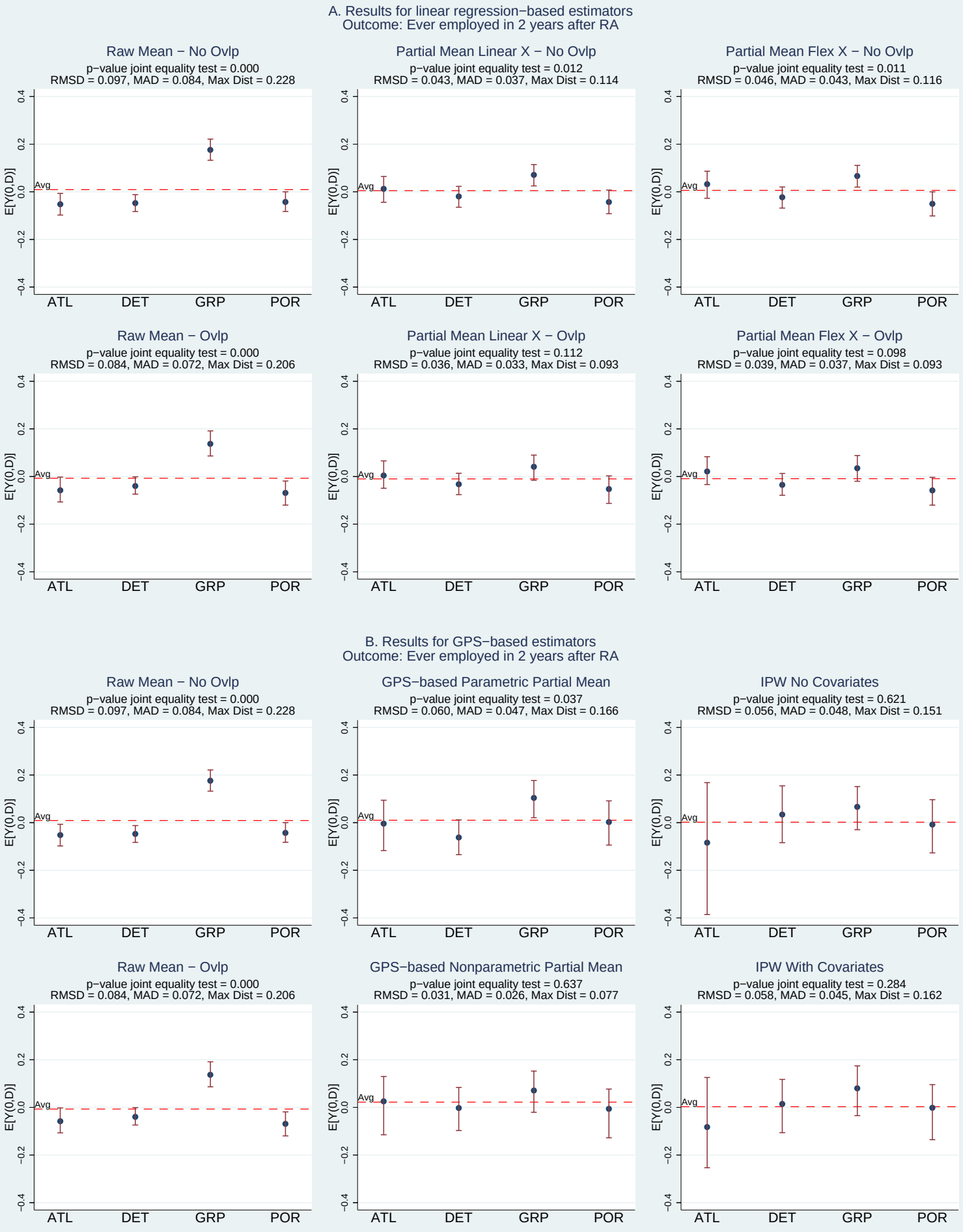
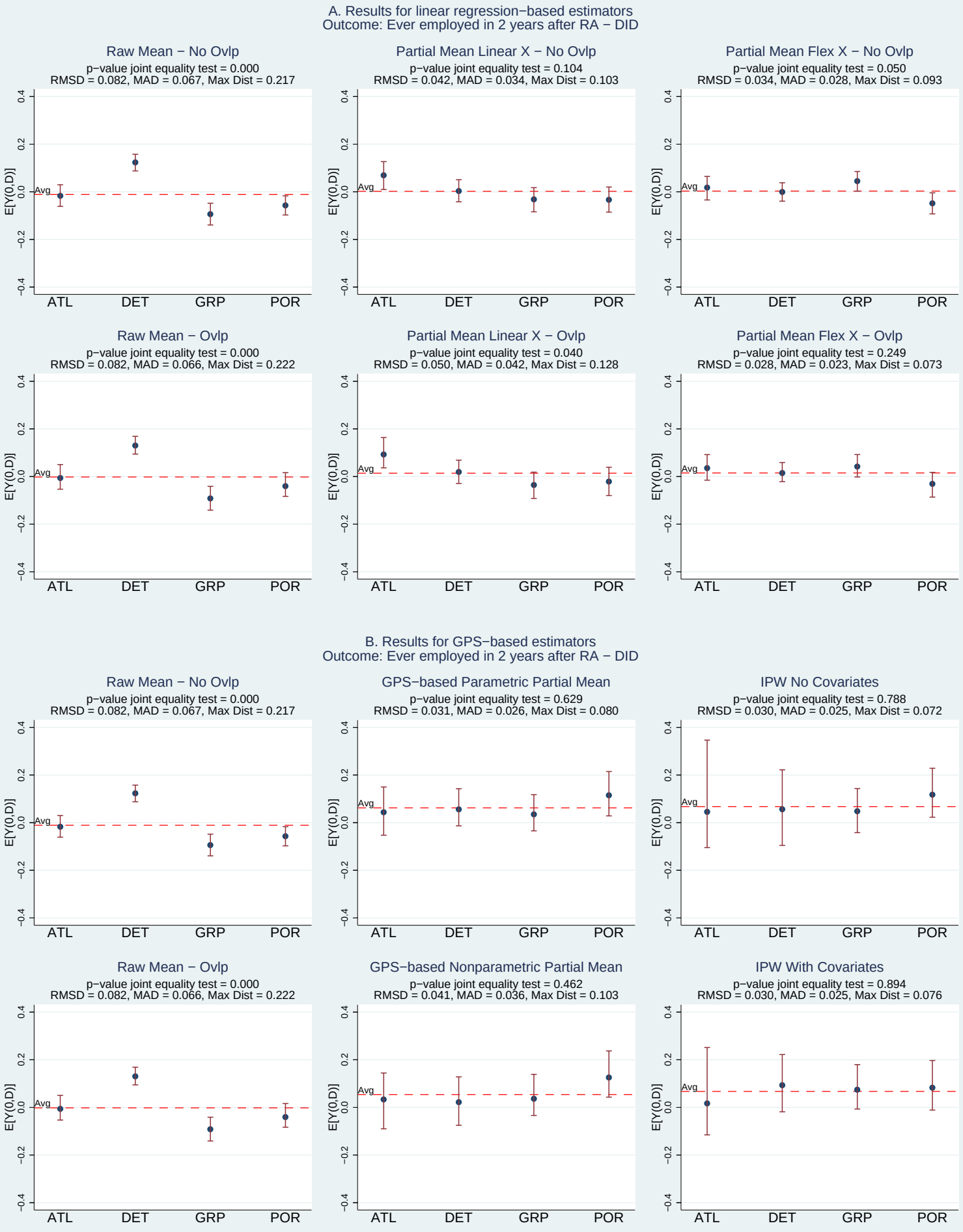


Figure 7. Estimated mean outcome in differences − 4 Sites



Appendix Table A1. Balancing of covariates based on joint equality of means tests across all sites - 5 sites

Variable	P-value joint equality of means test across all sites		Standardized means by site: Raw means					Means using GPS IPW				
	Raw	GPS IPW	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV
Black	0.000	0.005	0.94	0.81	-0.14	-0.57	-0.62	-0.02	0.17	0.32	0.40	0.38
Age 30-39 years old	0.000	0.335	0.21	-0.11	-0.22	-0.01	0.10	-0.13	-0.10	0.05	0.00	-0.07
Age 40+ years old	0.000	0.023	0.08	0.01	-0.09	-0.10	0.06	0.17	-0.16	-0.02	-0.02	0.01
Became mother as a teenager	0.000	0.081	0.09	0.08	0.21	-0.14	-0.12	-0.10	0.27	0.06	0.02	0.16
Never married	0.000	0.369	0.20	0.34	0.12	-0.05	-0.37	0.01	0.10	0.17	0.19	0.02
Any child 0-5 years old	0.000	0.093	-0.40	0.09	0.17	0.19	-0.07	0.13	0.15	-0.07	-0.03	-0.06
Any child 6-12 years old	0.000	0.059	0.32	-0.14	-0.24	-0.05	0.09	-0.07	-0.17	0.00	-0.09	0.19
2 children in household	0.004	0.854	0.03	-0.06	0.07	0.01	-0.01	0.07	0.04	-0.02	-0.03	0.07
3+ children in household	0.000	0.596	0.08	-0.01	-0.19	0.06	0.03	-0.06	-0.07	0.05	-0.02	-0.08
10th grade	0.000	0.916	0.01	0.03	-0.01	0.10	-0.08	-0.07	0.00	0.01	0.03	0.00
11th grade	0.000	0.841	-0.07	0.12	-0.01	0.03	-0.07	0.03	0.03	0.04	0.02	0.19
Grade 12 or higher	0.000	0.480	0.09	-0.06	0.02	-0.16	0.09	0.07	0.04	-0.07	0.00	-0.15
Highest degree = HS/GED	0.000	0.362	-0.02	-0.12	0.01	-0.01	0.10	-0.02	0.06	-0.11	-0.13	-0.15
Lives public/subss house	0.000	0.017	0.95	-0.34	-0.12	0.21	-0.29	-0.07	0.29	0.12	0.00	0.27
1-2 moves in past 2 years	0.000	0.833	-0.03	-0.04	0.02	-0.06	0.07	0.06	-0.06	0.00	0.06	0.00
3+ moves in past 2 years	0.000	0.772	-0.25	-0.25	0.20	0.14	0.12	-0.09	0.05	-0.13	-0.10	-0.03
On welfare < 2 years	0.000	0.095	-0.16	-0.22	0.10	-0.03	0.21	-0.14	0.09	-0.14	-0.19	-0.02
On welfare for 2-5 years	0.000	0.924	-0.08	-0.08	0.04	0.14	-0.01	0.02	0.07	0.05	-0.01	0.04
On welfare 5-10 years	0.000	0.038	0.09	0.09	-0.07	0.06	-0.11	0.32	-0.10	0.00	0.09	0.07
On welfare Q1 before RA	0.000	0.668	0.38	0.22	-0.13	-0.08	-0.23	0.15	0.17	0.22	0.17	0.16
On welfare Q2 before RA	0.000	0.224	0.48	0.32	-0.04	0.06	-0.48	0.21	0.21	0.26	0.28	0.36
On welfare Q3 before RA	0.000	0.010	0.35	0.34	0.00	0.08	-0.47	0.41	0.14	0.25	0.28	0.36
On welfare Q4 before RA	0.000	0.035	0.18	0.37	0.05	0.09	-0.44	0.38	0.11	0.22	0.25	0.32
On welfare Q5 before RA	0.000	0.012	0.15	0.40	0.06	0.06	-0.42	0.37	0.06	0.19	0.23	0.35
On welfare Q6 before RA	0.000	0.050	0.15	0.41	0.04	0.04	-0.41	0.38	0.01	0.23	0.25	0.26
On welfare Q7 before RA	0.000	0.148	0.15	0.41	-0.01	0.03	-0.39	0.34	0.02	0.23	0.26	0.27
Rec. FS in Q1 before RA	0.000	0.000	0.39	0.32	0.07	0.10	-0.51	0.40	0.19	0.25	0.25	0.19
Rec. FS in Q2 before RA	0.000	0.149	0.50	0.37	0.08	0.20	-0.67	0.44	0.27	0.30	0.32	0.37
Rec. FS in Q3 before RA	0.000	0.154	0.45	0.38	0.06	0.21	-0.65	0.43	0.17	0.30	0.33	0.32
Rec. FS in Q4 before RA	0.000	0.353	0.35	0.41	0.11	0.21	-0.65	0.40	0.19	0.29	0.32	0.31
Rec. FS in Q5 before RA	0.000	0.150	0.31	0.42	0.09	0.19	-0.61	0.38	0.09	0.29	0.32	0.34
Rec. FS in Q6 before RA	0.000	0.082	0.29	0.42	0.07	0.20	-0.60	0.42	0.08	0.30	0.31	0.33
Rec. FS in Q7 before RA	0.000	0.274	0.29	0.42	0.04	0.18	-0.57	0.31	0.09	0.27	0.31	0.22
Employed Q1 before RA	0.000	0.395	-0.08	-0.09	0.17	0.03	0.00	-0.06	-0.11	-0.02	-0.07	-0.17
Employed Q2 before RA	0.000	0.585	-0.13	-0.12	0.16	0.05	0.04	-0.08	-0.08	-0.06	-0.10	-0.18
Employed Q3 before RA	0.000	0.976	-0.11	-0.14	0.14	0.04	0.06	-0.05	-0.05	-0.07	-0.10	-0.07
Employed Q4 before RA	0.000	0.625	-0.05	-0.15	0.13	-0.01	0.08	-0.04	-0.05	-0.01	-0.11	-0.09
Employed Q5 before RA	0.000	0.910	-0.01	-0.17	0.14	-0.02	0.07	-0.05	-0.08	-0.01	-0.06	-0.06
Employed Q6 before RA	0.000	0.107	0.01	-0.19	0.18	-0.02	0.06	-0.04	-0.06	-0.02	-0.12	-0.21
Employed Q7 before RA	0.000	0.382	0.04	-0.20	0.20	-0.01	0.03	-0.01	-0.11	-0.02	-0.13	-0.09
Employed Q8 before RA	0.000	0.067	0.04	-0.23	0.24	-0.01	0.04	0.10	-0.08	-0.05	-0.18	0.00
Emply at RA (self reported)	0.000	0.042	-0.08	-0.10	0.10	-0.05	0.09	-0.13	-0.15	-0.02	-0.07	0.03
Ever wrkd FT 6+ mths same job	0.000	0.861	0.13	-0.41	-0.03	0.23	0.11	-0.08	-0.02	-0.05	-0.11	-0.16
Earnings Q1 before RA	0.000	0.596	-0.10	-0.12	0.04	0.01	0.10	0.17	-0.08	0.00	-0.09	-0.12
Earnings Q2 before RA	0.000	0.480	-0.15	-0.15	0.07	-0.02	0.16	0.16	-0.06	-0.04	-0.14	-0.13
Earnings Q3 before RA	0.000	0.512	-0.14	-0.16	0.06	-0.05	0.19	0.11	-0.05	-0.09	-0.14	-0.15
Earnings Q4 before RA	0.000	0.531	-0.06	-0.19	0.03	-0.05	0.18	0.13	0.02	-0.06	-0.10	-0.13
Earnings Q5 before RA	0.000	0.534	-0.02	-0.18	0.02	-0.06	0.17	0.10	-0.12	-0.05	-0.12	-0.11
Earnings Q6 before RA	0.000	0.530	0.02	-0.19	0.02	-0.05	0.14	0.04	-0.07	0.00	-0.12	-0.13
Earnings Q7 before RA	0.000	0.491	0.06	-0.20	0.03	-0.06	0.13	0.02	-0.04	-0.01	-0.12	-0.11
Earnings Q8 before RA	0.000	0.435	0.06	-0.20	0.03	-0.05	0.13	0.10	-0.05	-0.04	-0.13	-0.05
Any earns yr before RA (slf-rep)	0.000	0.420	-0.22	-0.28	0.27	0.07	0.14	-0.28	-0.23	-0.15	-0.17	-0.10
Emp/pop gr. rate 2 yrs before RA	0.000	0.000	-0.36	0.76	0.08	0.85	-0.94	0.15	-0.01	0.20	0.68	-0.68

Notes: Variables have been standardized to mean zero and standard deviation 1 (before imposing overlap).
GPS-based balancing tests are applied only to observations that satisfy the overlap condition.

Appendix Table A2. Balancing of covariates based on difference of means tests in each site versus all other sites pooled - 5 sites

Variable	Raw difference of means (standardized)					Difference of means after blocking on GPS (std)				
	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV
Black	1.10***	1.04***	-0.17***	-0.70***	-0.89***	0.24***	0.21***	0.11***	0.31***	0.08
Age 30-39 years old	0.24***	-0.14***	-0.26***	-0.02	0.14***	0.01	-0.04	0.04	-0.01	-0.05
Age 40+ years old	0.09***	0.01	-0.10***	-0.12***	0.09***	0.03	-0.05	-0.01	-0.02	-0.10
Became mother as a teenager	0.11***	0.10***	0.24***	-0.17***	-0.17***	0.01	0.05	0.02	-0.03	0.20***
Never married	0.24***	0.44***	0.14***	-0.07**	-0.53***	0.07	0.07*	0.02	0.08	0.00
Any child 0-5 years old	-0.47***	0.11***	0.19***	0.24***	-0.09***	-0.10	0.04	-0.05	0.01	0.01
Any child 6-12 years old	0.37***	-0.18***	-0.28***	-0.06**	0.13***	0.11*	-0.08*	0.00	0.00	0.07
2 children in household	0.03	-0.08***	0.08***	0.02	-0.01	0.06	0.02	0.01	0.00	0.09
3+ children in household	0.10***	-0.01	-0.23***	0.07**	0.04*	-0.03	-0.04	0.02	0.03	-0.12*
10th grade	0.01	0.03	-0.01	0.13***	-0.12***	-0.04	-0.02	-0.01	-0.06	-0.05
11th grade	-0.09***	0.16***	-0.01	0.04	-0.10***	-0.11**	0.01	0.03	0.00	0.00
Grade 12 or higher	0.10***	-0.08***	0.03	-0.20***	0.13***	0.11*	0.03	-0.04	0.10*	0.06
Highest degree = HS/GED	-0.03	-0.15***	0.01	-0.01	0.14***	0.03	-0.03	-0.07*	-0.04	0.01
Lives public/subss house	1.12***	-0.43***	-0.14***	0.26***	-0.42***	0.07	-0.12**	0.01	-0.05	0.17*
1-2 moves in past 2 years	-0.03	-0.05**	0.02	-0.07***	0.10***	0.05	0.05	0.05	0.07	0.06
3+ moves in past 2 years	-0.29***	-0.33***	0.24***	0.17***	0.18***	-0.05	-0.08**	-0.04	0.05	-0.01
On welfare < 2 years	-0.19***	-0.29***	0.12***	-0.03	0.30***	0.02	0.08*	-0.01	-0.02	0.05
On welfare for 2-5 years	-0.09***	-0.10***	0.05*	0.17***	-0.02	-0.06	-0.02	0.01	-0.03	-0.06
On welfare 5-10 years	0.11***	0.11***	-0.09***	0.07**	-0.15***	0.09	-0.07	-0.02	0.05	-0.01
On welfare Q1 before RA	0.45***	0.28***	-0.16***	-0.10***	-0.33***	-0.03	-0.04	0.02	-0.11**	-0.01
On welfare Q2 before RA	0.56***	0.41***	-0.04	0.08***	-0.69***	-0.06	-0.04	-0.02	-0.04	0.03
On welfare Q3 before RA	0.42***	0.44***	0.01	0.09***	-0.67***	0.00	-0.03	-0.02	-0.02	0.08
On welfare Q4 before RA	0.21***	0.48***	0.06**	0.11***	-0.63***	-0.05	-0.05	-0.03	-0.05	0.09
On welfare Q5 before RA	0.18***	0.51***	0.07**	0.07***	-0.61***	-0.02	-0.03	-0.01	-0.04	0.10
On welfare Q6 before RA	0.17***	0.52***	0.05*	0.05*	-0.59***	0.01	-0.04	0.03	0.01	0.11*
On welfare Q7 before RA	0.18***	0.53***	-0.01	0.04	-0.56***	0.02	-0.03	0.04	0.00	0.11
Rec. FS in Q1 before RA	0.46***	0.40***	0.08***	0.13***	-0.74***	0.09**	0.01	0.00	-0.08*	-0.04
Rec. FS in Q2 before RA	0.59***	0.47***	0.09***	0.25***	-0.97***	0.06	-0.01	-0.02	-0.08*	0.02
Rec. FS in Q3 before RA	0.53***	0.49***	0.07**	0.25***	-0.93***	0.10**	-0.02	-0.01	-0.05	0.01
Rec. FS in Q4 before RA	0.41***	0.53***	0.13***	0.26***	-0.93***	0.06	0.01	0.01	-0.06	0.01
Rec. FS in Q5 before RA	0.36***	0.53***	0.10***	0.24***	-0.88***	0.04	-0.02	0.03	-0.03	0.06
Rec. FS in Q6 before RA	0.34***	0.53***	0.09***	0.24***	-0.86***	0.08	-0.02	0.04	0.00	0.04
Rec. FS in Q7 before RA	0.34***	0.54***	0.05*	0.22***	-0.82***	0.07	-0.02	0.03	0.01	0.03
Employed Q1 before RA	-0.10***	-0.11***	0.20***	0.04	0.00	-0.04	-0.02	0.04	0.00	0.02
Employed Q2 before RA	-0.15***	-0.15***	0.19***	0.06**	0.06**	-0.03	-0.03	0.00	-0.01	0.08
Employed Q3 before RA	-0.12***	-0.18***	0.16***	0.04	0.09***	0.01	-0.04	-0.01	-0.01	0.13
Employed Q4 before RA	-0.06**	-0.20***	0.15***	-0.02	0.12***	0.01	-0.05	0.05	-0.03	0.08
Employed Q5 before RA	-0.01	-0.22***	0.16***	-0.02	0.10***	0.01	-0.02	0.06	-0.03	0.14*
Employed Q6 before RA	0.01	-0.24***	0.21***	-0.03	0.08***	0.03	-0.05	0.03	-0.07	0.08
Employed Q7 before RA	0.04	-0.25***	0.24***	-0.02	0.05**	0.02	-0.03	0.03	-0.05	0.04
Employed Q8 before RA	0.04	-0.29***	0.28***	-0.01	0.06**	0.08	-0.04	0.00	-0.10**	0.13
Emply at RA (self reported)	-0.10***	-0.13***	0.12***	-0.06**	0.13***	-0.06	-0.02	0.03	0.01	0.07
Ever wrkd FT 6+ mths same job	0.15***	-0.52***	-0.04	0.28***	0.16***	0.01	-0.09**	-0.01	0.01	-0.05
Earnings Q1 before RA	-0.11***	-0.15***	0.05	0.01	0.15***	0.00	-0.02	0.08	0.00	0.05
Earnings Q2 before RA	-0.17***	-0.20***	0.08***	-0.03	0.23***	0.03	-0.01	0.06	0.02	0.09
Earnings Q3 before RA	-0.17***	-0.21***	0.07**	-0.06***	0.27***	0.04	-0.01	0.03	0.03	0.04
Earnings Q4 before RA	-0.07***	-0.24***	0.03	-0.06***	0.26***	0.05	-0.02	0.04	0.02	-0.02
Earnings Q5 before RA	-0.03	-0.23***	0.02	-0.07***	0.24***	0.05	-0.04	0.07	0.00	0.04
Earnings Q6 before RA	0.03	-0.24***	0.03	-0.06***	0.21***	0.06	-0.02	0.07	-0.01	0.03
Earnings Q7 before RA	0.07**	-0.25***	0.04	-0.07***	0.19***	0.03	0.01	0.06	0.00	0.01
Earnings Q8 before RA	0.07**	-0.26***	0.04	-0.06**	0.19***	0.05	0.01	0.05	0.00	0.02
Any earns yr before RA (slf-rep)	-0.26***	-0.36***	0.31***	0.08***	0.20***	-0.05	-0.08**	-0.05	0.02	0.10
Emp/pop gr. rate 2 yrs before RA	-0.42***	0.98***	0.09***	1.05***	-1.34***	-0.04	0.07**	-0.02	0.47***	-0.94***

Notes: * Significant at 10%; ** Significant at 5%; *** Significant at 1%.

Variables have been standardized to mean zero and standard deviation 1 (before imposing overlap).

GPS-based balancing tests are applied only to observations that satisfy the overlap condition.

Appendix Table A3. Balancing of covariates based on joint equality of means tests across all sites - 4 sites

Variable	P-value joint equality of means test across all sites		Standardized means by site							
			Raw means				Means using GPS IPW			
	Raw	GPS IPW	ATL	DET	GRP	POR	ATL	DET	GRP	POR
Black	0.000	0.968	0.69	0.56	-0.42	-0.86	0.10	0.14	0.15	0.16
Age 30-39 years old	0.000	0.873	0.25	-0.07	-0.18	0.03	-0.01	0.03	0.07	0.02
Age 40+ years old	0.000	0.206	0.11	0.03	-0.06	-0.07	-0.11	-0.09	0.00	0.01
Became mother as a teenager	0.000	0.197	0.04	0.03	0.15	-0.19	-0.09	0.14	0.02	-0.02
Never married	0.000	0.912	0.04	0.18	-0.04	-0.22	0.04	0.00	0.06	0.02
Any child 0-5 years old	0.000	0.076	-0.43	0.06	0.14	0.17	0.14	0.05	-0.08	-0.07
Any child 6-12 years old	0.000	0.630	0.36	-0.10	-0.20	-0.01	-0.07	-0.04	0.04	-0.03
2 children in household	0.002	0.891	0.02	-0.07	0.06	0.01	0.01	0.01	-0.04	-0.02
3+ children in household	0.000	0.596	0.09	0.00	-0.18	0.07	-0.06	0.01	0.06	-0.01
10th grade	0.013	0.692	-0.03	-0.01	-0.05	0.07	-0.11	0.00	-0.02	-0.02
11th grade	0.000	0.644	-0.10	0.09	-0.04	0.00	0.05	-0.06	0.01	-0.02
Grade 12 or higher	0.000	0.572	0.13	-0.02	0.06	-0.12	0.09	0.07	-0.02	0.06
Highest degree = HS/GED	0.001	0.788	0.02	-0.08	0.05	0.04	-0.01	-0.01	-0.06	-0.09
Lives public/subss house	0.000	0.003	0.77	-0.43	-0.23	0.08	-0.08	0.18	0.04	-0.13
1-2 moves in past 2 years	0.161	0.759	0.00	-0.01	0.05	-0.03	0.07	0.09	0.03	0.11
3+ moves in past 2 years	0.000	0.336	-0.21	-0.21	0.27	0.20	-0.17	0.01	-0.10	-0.06
On welfare < 2 years	0.000	0.037	-0.07	-0.14	0.20	0.06	-0.04	0.12	-0.07	-0.11
On welfare for 2-5 years	0.000	0.821	-0.09	-0.08	0.04	0.13	0.03	0.00	0.04	-0.02
On welfare 5-10 years	0.000	0.277	0.04	0.04	-0.12	0.01	0.18	-0.06	-0.03	0.04
On welfare Q1 before RA	0.000	0.307	0.31	0.13	-0.26	-0.20	-0.04	0.09	0.15	0.07
On welfare Q2 before RA	0.000	0.554	0.31	0.13	-0.28	-0.17	-0.11	0.04	0.07	0.07
On welfare Q3 before RA	0.000	0.303	0.17	0.16	-0.22	-0.14	0.17	0.01	0.07	0.07
On welfare Q4 before RA	0.000	0.097	-0.01	0.20	-0.15	-0.11	0.17	-0.05	0.05	0.04
On welfare Q5 before RA	0.000	0.102	-0.03	0.23	-0.13	-0.13	0.18	-0.04	0.02	0.03
On welfare Q6 before RA	0.000	0.031	-0.03	0.24	-0.14	-0.15	0.22	-0.06	0.07	0.06
On welfare Q7 before RA	0.000	0.094	-0.02	0.25	-0.18	-0.14	0.21	-0.04	0.07	0.08
Rec. FS in Q1 before RA	0.000	0.035	0.22	0.12	-0.20	-0.16	0.19	0.12	0.08	0.05
Rec. FS in Q2 before RA	0.000	0.816	0.27	0.10	-0.27	-0.11	0.12	0.06	0.06	0.03
Rec. FS in Q3 before RA	0.000	0.753	0.21	0.12	-0.27	-0.09	0.11	0.01	0.06	0.04
Rec. FS in Q4 before RA	0.000	0.801	0.08	0.15	-0.20	-0.08	0.10	0.01	0.05	0.02
Rec. FS in Q5 before RA	0.000	0.597	0.05	0.17	-0.20	-0.08	0.08	-0.03	0.06	0.04
Rec. FS in Q6 before RA	0.000	0.155	0.04	0.17	-0.21	-0.07	0.14	-0.06	0.09	0.04
Rec. FS in Q7 before RA	0.000	0.524	0.04	0.19	-0.22	-0.08	0.08	-0.04	0.07	0.06
Employed Q1 before RA	0.000	0.870	-0.08	-0.09	0.17	0.03	-0.09	-0.04	-0.03	-0.07
Employed Q2 before RA	0.000	0.866	-0.11	-0.10	0.18	0.06	-0.01	-0.02	-0.05	-0.09
Employed Q3 before RA	0.000	0.809	-0.08	-0.11	0.17	0.06	0.05	-0.05	-0.05	-0.08
Employed Q4 before RA	0.000	0.540	-0.02	-0.12	0.17	0.02	0.05	-0.02	0.01	-0.08
Employed Q5 before RA	0.000	0.973	0.02	-0.14	0.17	0.02	0.02	0.00	0.01	-0.03
Employed Q6 before RA	0.000	0.602	0.04	-0.16	0.21	0.00	-0.05	0.00	-0.01	-0.09
Employed Q7 before RA	0.000	0.528	0.05	-0.19	0.22	0.00	-0.05	-0.02	-0.02	-0.11
Employed Q8 before RA	0.000	0.097	0.05	-0.21	0.26	0.01	0.03	-0.03	-0.04	-0.16
Emply at RA (self reported)	0.000	0.157	-0.05	-0.07	0.15	-0.01	-0.12	-0.08	-0.01	-0.03
Ever wrkd FT 6+ mths same job	0.000	0.926	0.17	-0.36	0.01	0.27	0.01	0.00	-0.02	-0.06
Earnings Q1 before RA	0.000	0.851	-0.06	-0.08	0.10	0.06	0.03	-0.02	0.00	-0.05
Earnings Q2 before RA	0.000	0.422	-0.09	-0.10	0.17	0.06	0.14	0.02	-0.02	-0.08
Earnings Q3 before RA	0.000	0.573	-0.08	-0.10	0.17	0.04	0.13	0.01	-0.04	-0.06
Earnings Q4 before RA	0.000	0.657	0.02	-0.13	0.13	0.03	0.13	0.05	-0.03	-0.02
Earnings Q5 before RA	0.000	0.713	0.06	-0.13	0.11	0.02	0.08	-0.02	-0.02	-0.05
Earnings Q6 before RA	0.000	0.758	0.10	-0.14	0.10	0.01	-0.03	0.02	0.03	-0.05
Earnings Q7 before RA	0.000	0.596	0.13	-0.16	0.10	0.00	-0.03	0.05	0.03	-0.06
Earnings Q8 before RA	0.000	0.610	0.13	-0.17	0.10	0.01	0.01	0.02	0.00	-0.07
Any earns yr before RA (slf-rep)	0.000	0.936	-0.17	-0.22	0.33	0.13	-0.17	-0.12	-0.11	-0.11
Emp/pop gr. rate 2 yrs before RA	0.000	0.000	-0.83	0.38	-0.35	0.48	0.17	-0.18	-0.13	0.29

Notes: Variables have been standardized to mean zero and standard deviation 1 (before imposing overlap).
GPS-based balancing tests are applied only to observations that satisfy the overlap condition.

Appendix Table A4. Balancing of covariates based on difference of means tests in each site versus all other sites pooled - 4 sites

Variable	Raw difference of means (standardized)				Difference of means after blocking on GPS (std)			
	ATL	DET	GRP	POR	ATL	DET	GRP	POR
Black	0.87***	0.81***	-0.53***	-1.17***	0.26***	0.12***	0.02	0.07***
Age 30-39 years old	0.32***	-0.10***	-0.23***	0.04	0.00	-0.05	0.02	0.00
Age 40+ years old	0.14***	0.05*	-0.08***	-0.10***	0.00	-0.01	-0.03	-0.02
Became mother as a teenager	0.05*	0.04*	0.19***	-0.26***	0.02	0.06	0.02	-0.06
Never married	0.05*	0.27***	-0.05	-0.30***	0.07	0.04	0.02	-0.01
Any child 0-5 years old	-0.55***	0.08***	0.17***	0.23***	-0.08	0.01	0.00	0.04
Any child 6-12 years old	0.45***	-0.14***	-0.25***	-0.01	0.09	-0.03	-0.03	0.03
2 children in household	0.03	-0.10***	0.08**	0.01	0.04	0.02	-0.01	-0.01
3+ children in household	0.12***	0.00	-0.23***	0.09***	-0.02	-0.01	0.02	0.05
10th grade	-0.03	-0.01	-0.06**	0.09***	-0.07	0.00	0.00	-0.05
11th grade	-0.13***	0.14***	-0.05*	0.00	-0.07	0.01	0.02	0.00
Grade 12 or higher	0.16***	-0.03	0.08***	-0.17***	0.12*	0.00	-0.01	0.08
Highest degree = HS/GED	0.02	-0.11***	0.06**	0.05*	0.05	-0.04	-0.04	-0.03
Lives public/subss house	0.97***	-0.62***	-0.29***	0.10***	0.00	-0.14***	0.00	-0.07
1-2 moves in past 2 years	0.00	-0.02	0.06**	-0.04	0.00	0.04	0.06	0.07
3+ moves in past 2 years	-0.26***	-0.31***	0.34***	0.27***	0.02	-0.04	-0.02	0.08
On welfare < 2 years	-0.09***	-0.20***	0.25***	0.09***	0.05	0.07	0.01	0.01
On welfare for 2-5 years	-0.11***	-0.12***	0.05	0.18***	-0.08	-0.01	0.03	-0.01
On welfare 5-10 years	0.05*	0.06**	-0.15***	0.01	0.10*	-0.07*	-0.02	0.04
On welfare Q1 before RA	0.40***	0.19***	-0.33***	-0.27***	-0.05	-0.01	0.01	-0.13**
On welfare Q2 before RA	0.39***	0.18***	-0.36***	-0.23***	-0.09	-0.03	-0.03	-0.08
On welfare Q3 before RA	0.21***	0.23***	-0.28***	-0.19***	-0.05	-0.03	-0.02	-0.05
On welfare Q4 before RA	-0.01	0.29***	-0.19***	-0.15***	-0.03	-0.03	-0.04	-0.08
On welfare Q5 before RA	-0.04	0.33***	-0.17***	-0.18***	-0.01	-0.02	-0.04	-0.09
On welfare Q6 before RA	-0.04	0.35***	-0.18***	-0.20***	0.00	-0.02	0.00	-0.05
On welfare Q7 before RA	-0.02	0.37***	-0.23***	-0.19***	0.01	-0.02	0.01	-0.04
Rec. FS in Q1 before RA	0.28***	0.18***	-0.26***	-0.21***	-0.01	0.02	-0.01	-0.13**
Rec. FS in Q2 before RA	0.34***	0.14***	-0.34***	-0.15***	-0.06	-0.01	-0.01	-0.15**
Rec. FS in Q3 before RA	0.26***	0.17***	-0.34***	-0.12***	-0.02	-0.03	-0.01	-0.10*
Rec. FS in Q4 before RA	0.10***	0.22***	-0.25***	-0.12***	-0.03	0.00	0.00	-0.12**
Rec. FS in Q5 before RA	0.06**	0.25***	-0.26***	-0.11***	-0.04	-0.04	0.00	-0.09*
Rec. FS in Q6 before RA	0.05	0.25***	-0.26***	-0.09***	-0.03	-0.05	0.02	-0.06
Rec. FS in Q7 before RA	0.05*	0.27***	-0.28***	-0.10***	-0.01	-0.04	0.02	-0.05
Employed Q1 before RA	-0.11***	-0.13***	0.22***	0.04	0.03	-0.01	0.02	0.00
Employed Q2 before RA	-0.14***	-0.15***	0.23***	0.09***	-0.01	-0.02	0.00	0.00
Employed Q3 before RA	-0.10***	-0.17***	0.21***	0.09***	0.06	-0.03	0.00	0.00
Employed Q4 before RA	-0.02	-0.18***	0.21***	0.03	0.06	-0.01	0.04	-0.01
Employed Q5 before RA	0.03	-0.21***	0.22***	0.02	0.04	0.00	0.04	-0.01
Employed Q6 before RA	0.05	-0.24***	0.26***	0.00	0.03	-0.03	0.01	-0.05
Employed Q7 before RA	0.07**	-0.27***	0.28***	0.00	0.01	-0.04	0.01	-0.04
Employed Q8 before RA	0.07**	-0.31***	0.32***	0.01	0.07	-0.05	0.00	-0.08*
Emply at RA (self reported)	-0.06**	-0.09***	0.19***	-0.01	0.01	-0.01	0.02	0.03
Ever wrkd FT 6+ mths same job	0.22***	-0.52***	0.02	0.37***	0.05	-0.05	0.01	0.04
Earnings Q1 before RA	-0.08**	-0.12***	0.13***	0.09***	0.03	-0.01	0.02	0.02
Earnings Q2 before RA	-0.12***	-0.14***	0.21***	0.08***	0.05	0.01	0.00	0.04
Earnings Q3 before RA	-0.10***	-0.15***	0.22***	0.06**	0.09	0.00	-0.01	0.05
Earnings Q4 before RA	0.02	-0.19***	0.16***	0.05	0.08	0.02	0.00	0.05
Earnings Q5 before RA	0.08**	-0.19***	0.14***	0.02	0.06	-0.01	0.01	0.02
Earnings Q6 before RA	0.13***	-0.21***	0.12***	0.02	0.05	-0.01	0.00	0.00
Earnings Q7 before RA	0.17***	-0.23***	0.13***	0.00	0.02	0.01	0.00	0.00
Earnings Q8 before RA	0.17***	-0.24***	0.13***	0.02	0.06	-0.01	0.00	0.01
Any earns yr before RA (slf-rep)	-0.21***	-0.33***	0.42***	0.18***	-0.01	-0.05	-0.02	0.06
Emp/pop gr. rate 2 yrs before RA	-1.05***	0.56***	-0.45***	0.66***	-0.04	0.03	-0.08***	0.34***

Notes: * Significant at 10%; ** Significant at 5%; *** Significant at 1%.

Variables have been standardized to mean zero and standard deviation 1 (before imposing overlap).

GPS-based balancing tests are applied only to observations that satisfy the overlap condition.

Appendix Table A5. Estimated average employment rate in two years after random assignment - 5 sites

Estimator	ATL	DET	GRP	POR	RIV
A. Outcome in levels					
Raw Mean - No Ovlp	0.04	0.05	0.27	0.05	-0.22
	[-0.01,0.09]	[0.01,0.09]	[0.22,0.31]	[0.01,0.10]	[-0.25,-0.19]
Raw Mean - Ovlp	0.04	0.06	0.24	0.03	-0.29
	[-0.02,0.09]	[0.02,0.09]	[0.18,0.30]	[-0.02,0.08]	[-0.35,-0.24]
Linear regression-based					
Partial Mean Linear X - No Ovlp	0.08	0.07	0.15	0.03	-0.18
	[0.01,0.13]	[0.02,0.12]	[0.11,0.20]	[-0.02,0.08]	[-0.22,-0.14]
Partial Mean Linear X - Ovlp	0.08	0.05	0.12	0.01	-0.19
	[0.01,0.13]	[0.00,0.10]	[0.07,0.18]	[-0.04,0.07]	[-0.26,-0.13]
Partial Mean Flex X - No Ovlp	0.09	0.07	0.15	0.03	-0.18
	[0.03,0.15]	[0.02,0.12]	[0.10,0.20]	[-0.03,0.08]	[-0.23,-0.14]
Partial Mean Flex X - Ovlp	0.09	0.05	0.12	0.01	-0.19
	[0.02,0.15]	[0.00,0.10]	[0.07,0.17]	[-0.04,0.07]	[-0.26,-0.13]
GPS-based (imposing Ovlp)					
Parametric Partial Mean	0.08	0.01	0.13	0.09	-0.20
	[-0.02,0.17]	[-0.07,0.09]	[0.06,0.22]	[-0.02,0.19]	[-0.41,-0.05]
Nonparametric Partial Mean	0.13	0.08	0.12	0.09	-0.27
	[-0.01,0.24]	[-0.04,0.19]	[0.03,0.21]	[-0.08,0.19]	[-0.50,-0.09]
IPW No Covariates	0.04	0.14	0.13	0.08	-0.30
	[-0.38,0.24]	[-0.02,0.33]	[0.04,0.22]	[-0.04,0.19]	[-0.50,-0.09]
IPW With Covariates	0.07	0.10	0.12	0.07	-0.25
	[-0.19,0.22]	[-0.05,0.25]	[0.00,0.21]	[-0.07,0.19]	[-0.51,-0.10]
B. Outcome in differences (with respecto to years 1 and 2 before RA)					
Raw Estimator - No Ovlp	0.06	0.20	-0.02	0.02	-0.18
	[0.01,0.11]	[0.16,0.24]	[-0.06,0.03]	[-0.02,0.06]	[-0.21,-0.15]
Raw Estimator - Ovlp	0.07	0.21	-0.01	0.04	-0.15
	[0.02,0.13]	[0.17,0.25]	[-0.07,0.04]	[-0.01,0.09]	[-0.22,-0.10]
Linear regression-based					
Partial Mean Linear X - No Ovlp	0.10	0.07	0.02	0.01	-0.12
	[0.04,0.16]	[0.02,0.12]	[-0.03,0.08]	[-0.04,0.06]	[-0.16,-0.07]
Partial Mean Linear X - Ovlp	0.15	0.08	0.03	0.04	-0.07
	[0.08,0.22]	[0.03,0.14]	[-0.02,0.09]	[-0.02,0.10]	[-0.15,-0.01]
Partial Mean Flex X - No Ovlp	0.07	0.08	0.12	0.02	-0.16
	[0.02,0.13]	[0.03,0.12]	[0.08,0.17]	[-0.03,0.06]	[-0.20,-0.12]
Partial Mean Flex X - Ovlp	0.11	0.09	0.13	0.04	-0.13
	[0.04,0.15]	[0.05,0.14]	[0.08,0.17]	[-0.02,0.09]	[-0.20,-0.08]
GPS-based (imposing Ovlp)					
Parametric Partial Mean	0.09	0.14	0.09	0.21	-0.11
	[-0.01,0.18]	[0.05,0.23]	[0.02,0.18]	[0.09,0.31]	[-0.29,0.03]
Nonparametric Partial Mean	0.12	0.10	0.11	0.25	-0.13
	[-0.01,0.23]	[-0.04,0.22]	[0.03,0.25]	[0.13,0.41]	[-0.30,0.02]
IPW No Covariates	0.12	0.17	0.10	0.19	-0.14
	[-0.04,0.33]	[-0.07,0.47]	[0.02,0.19]	[0.09,0.28]	[-0.31,0.04]
IPW With Covariates	0.11	0.18	0.10	0.12	-0.01
	[-0.04,0.27]	[-0.01,0.38]	[0.02,0.21]	[0.02,0.26]	[-0.22,0.15]

Note: Bootstrap confidence intervals in brackets (based on 1,000 replications).

Appendix Table A6. Estimated average employment rate in two years after random assignment - 4 sites

Estimator	ATL	DET	GRP	POR
A. Outcome in levels				
Raw Mean - No Ovlp	-0.05	-0.05	0.18	-0.04
	[-0.10,-0.01]	[-0.08,-0.01]	[0.13,0.22]	[-0.08,0.00]
Raw Mean - Ovlp	-0.06	-0.04	0.14	-0.07
	[-0.11,0.00]	[-0.07,0.00]	[0.09,0.19]	[-0.12,-0.02]
Linear regression-based				
Partial Mean Linear X - No Ovlp	0.01	-0.02	0.07	-0.04
	[-0.04,0.06]	[-0.06,0.02]	[0.02,0.11]	[-0.09,0.01]
Partial Mean Linear X - Ovlp	0.00	-0.03	0.04	-0.05
	[-0.05,0.07]	[-0.08,0.01]	[-0.02,0.09]	[-0.11,0.00]
Partial Mean Flex X - No Ovlp	0.03	-0.02	0.07	-0.05
	[-0.03,0.09]	[-0.07,0.02]	[0.02,0.11]	[-0.10,0.00]
Partial Mean Flex X - Ovlp	0.02	-0.04	0.04	-0.06
	[-0.03,0.08]	[-0.08,0.01]	[-0.02,0.09]	[-0.12,0.00]
GPS-based (imposing Ovlp)				
Parametric Partial Mean	0.00	-0.06	0.10	0.00
	[-0.12,0.09]	[-0.13,0.01]	[0.02,0.18]	[-0.09,0.09]
Nonparametric Partial Mean	0.03	0.00	0.07	-0.01
	[-0.11,0.13]	[-0.10,0.08]	[-0.02,0.15]	[-0.13,0.08]
IPW No Covariates	-0.08	0.03	0.07	-0.01
	[-0.39,0.17]	[-0.08,0.15]	[-0.03,0.15]	[-0.13,0.10]
IPW With Covariates	-0.08	0.01	0.08	0.00
	[-0.25,0.13]	[-0.11,0.12]	[-0.03,0.17]	[-0.14,0.10]
B. Outcome in differences (with respecto to years 1 and 2 before RA)				
Raw Estimator - No Ovlp	-0.02	0.12	-0.09	-0.06
	[-0.06,0.03]	[0.09,0.16]	[-0.14,-0.05]	[-0.10,-0.02]
Raw Estimator - Ovlp	-0.01	0.13	-0.09	-0.04
	[-0.05,0.05]	[0.09,0.17]	[-0.14,-0.04]	[-0.08,0.02]
Linear regression-based				
Partial Mean Linear X - No Ovlp	0.07	0.00	-0.03	-0.03
	[0.01,0.13]	[-0.04,0.05]	[-0.08,0.02]	[-0.09,0.02]
Partial Mean Linear X - Ovlp	0.09	0.02	-0.04	-0.02
	[0.04,0.16]	[-0.03,0.07]	[-0.09,0.02]	[-0.08,0.04]
Partial Mean Flex X - No Ovlp	0.02	0.00	0.04	-0.05
	[-0.03,0.06]	[-0.04,0.04]	[0.00,0.09]	[-0.09,0.00]
Partial Mean Flex X - Ovlp	0.04	0.01	0.04	-0.03
	[-0.02,0.09]	[-0.02,0.06]	[0.00,0.09]	[-0.09,0.02]
GPS-based (imposing Ovlp)				
Parametric Partial Mean	0.04	0.06	0.03	0.11
	[-0.05,0.15]	[-0.01,0.14]	[-0.03,0.12]	[0.03,0.21]
Nonparametric Partial Mean	0.03	0.02	0.04	0.13
	[-0.09,0.14]	[-0.08,0.13]	[-0.03,0.14]	[0.04,0.24]
IPW No Covariates	0.05	0.06	0.05	0.12
	[-0.10,0.35]	[-0.10,0.22]	[-0.04,0.14]	[0.02,0.23]
IPW With Covariates	0.02	0.09	0.07	0.08
	[-0.12,0.25]	[-0.02,0.22]	[-0.01,0.18]	[-0.01,0.20]

Note: Bootstrap confidence intervals in brackets (based on 1,000 replications).