Evaluating Nonexperimental Estimators for Multiple Treatments: Evidence from a Randomized Experiment*

Carlos A. Flores† Oscar Mitnik‡

September 29, 2008

Preliminary and Incomplete - Comments Welcome

Abstract

This paper assesses nonexperimental estimators of mean effects of multiple or multivalued treatments by analyzing their effectiveness in adjusting for observable characteristics and eliminating differences in average outcomes among multiple populations. The data we use comes 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 seven locations were randomly assigned to a control group or to different training programs emphasizing either human capital development or labor force attachment. The prior literature evaluating the performance of nonexperimental methods has focused exclusively on binary treatments. Given the growing interest in evaluating programs in which the treatment is multivalued or there are more than one treatment it is important to learn about the performance of different estimators in this context. Among the estimators studied, we pay particularly attention to those based on the generalized propensity score or GPS, which equals the probability of receiving a particular treatment (or level of the treatment) conditional on covariates. In addition, we analyze the role of the GPS in identifying units across treatment groups that are comparable in terms of observable characteristics, and provide guidance for its use in practice.

*We thank Phil Robins and participants at the University of Miami Labor Discussion Group for useful comments and discussions. Bryan Mueller and Yongmin Zang provided excellent research assistance.

†Department of Economics, University of Miami. Email: caflores@miami.edu.

‡Department of Economics, University of Miami and IZA. Email: omitnik@miami.edu.

1 Introduction

Nonexperimental methods are widely used in economics and other disciplines to evaluate government programs and many types of interventions. And, in the absence of an experiment (or infeasibility of conducting one), nonexperimental methods are in many situations the only alternative. Among them, those based on selection on observables or unconfoundedness assumptions play an important role (e.g., Imbens, 2004, 2008; Heckman et al., 1999). Most of the focus on nonexperimental methods in the previous two decades has been on estimation of average treatment effects of a binary treatment or intervention on an outcome. In practice, however, individuals are usually exposed to different doses of the treatment or to more than one treatment. As a result, the focus has recently shifted to developing methods to evaluate such programs. This paper contributes to this literature by assessing the effectiveness of nonexperimental estimators of mean effects for multiple or multivalued treatments in adjusting for observable characteristics and eliminating differences in average outcomes among multiple populations. The data we use comes 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 seven locations were randomly assigned to a control group or to different training programs emphasizing either human capital development or labor force attachment.

Since the influential paper by Lalonde (1986) many studies have evaluated the performance of different nonexperimental methods (e.g., Heckman and Hotz, 1989; Friedlander and Robins, 1995; Heckman et al., 1997, 1998; Dehejia and Wahba, 1999, 2002; Michalopoulos et al., 2004; Smith and Todd, 2005; Dehejia, 2005). This literature has advanced our understanding of nonexperimental evaluations by specifying conditions under which nonexperimental estimators are more likely to replicate the outcome from a randomized experiment. One of the main conclusions is the importance of comparing "comparable" individuals. For instance, Heckman et al. (1997, 1998ab) 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 (i.e., the probability of receiving treatment conditional on covariates) to identify regions of the data where treatment and control units are comparable in terms of observed characteristics.

A common characteristic of the current literature evaluating nonexperimental estimators based on a selection-on-observables assumption is its focus on binary treatments: individuals either participate in a program or not. Recently, however, there has been a growing interest in evaluating programs or interventions in which the treatment is multivalued or there are more than one treatment (e.g.,Behrman et al., 2004; Frölich et al., 2004; Flores-Lagunes et al., 2007; Kluve et al., 2007; Mitnik, 2008), and on different methods to evaluate such programs (e.g., Imbens, 2000; Lechner, 2001; Hirano and Imbens, 2004; Cattaneo, 2007; Flores, 2007). Unfortunately, very little is known about the performance of alternative estimation techniques in terms of reducing the potential selection bias present in nonexperimental evaluation of multiple treatments. To our knowledge, ours is the first study to address this issue.

When the treatment is multivalued or there are more than one treatment we have more parameters of interest than the commonly used average treatment effect (or average treatment effect on the treated) in the binary-treatment case. For instance, one may be interested on pairwise comparisons (e.g., Lechner, 2001), or on finding the level of the treatment (or the particular treatment) that gives the highest average outcome (e.g., Flores, 2007). In this paper, we focus on estimators of what is some times called the dose-response function (although this may not be the most appropriate denomination for non-ordered multi-valued treatments). It gives the average potential outcome over all possible values of the treatment. In other words, it gives the expected potential outcome at all possible values of the treatment for someone randomly chosen from the population. Since in a nonexperimental evaluation the population is selected into different treatment levels, a major task for estimation of the dose-response function is finding individuals that are comparable simultaneously across all treatment levels.

The general approach to evaluate the performance of nonexperimental estimators in the binary-treatment case consists on using data from a randomized experiment and constructing a nonexperimental control group, for instance, from additional data sets (e.g., Lalonde, 1986) or from different locations (e.g., Friedlander and Robins, 1995). The different nonexperimental estimators are then used on the nonexperimental control group and the experimental treated group and, to asses the performance of the estimators, the results are compared against those from the experiment. One could also apply the estimators to the nonexperimental and experimental control groups, in which case the benchmark is obtaining a zero treatment effect. A special application of this general approach is Heckman et al. (1997), in which their nonexperimental control group consisted of individuals that (i) were eligible to the program being evaluated (the National Job Training Partnership Act, JTPA) but that did not apply, (ii) resided in the same narrowly defined area as the applicants; and, (iii) were administered the same survey as those in the experiment. As stressed in their paper, having a nonexperimental control group in the same local labor market as those receiving treatment, administering the same questionnaire and having detailed labor market history seem to be key for nonexperimental methods to work properly.

Extending the logic of the prior literature focused on binary treatments, an ideal data for the purpose of evaluating nonexperimental methods of multiple treatments would consist of an experiment in which units are randomized into s different treatments, with $s > 2$. In addition, for each of $s-1$ treatments there would be units that self-select into these same treatments but that are otherwise representative of the population in which the experiment took place (e.g., welfare recipients in a given area and time). These units would form the nonexperimental groups. The data would have to contain detailed information on all units (e.g., background characteristics and previous labor market history), and the same data gathering instrument would have been used for all units. In this case, we could take the nonexperimental groups plus one of the experimental groups and, applying (maybe alternative) nonexperimental methods we could compare the results to those from the actual experiment. Unfortunately, such a data is not available to the best of our knowledge, and we resort to a different strategy.

In this paper, we resort to the availability of several control groups in different sites of the NEWWS experiment to evaluate alternative nonexperimental estimators of multiple treatments. We use the methods to adjust for observable characteristics in order to eliminate difference in outcomes among members of the control groups in different sites. Relying on individuals from an experiment has the advantages that i) they may be relatively comparable (at least in this case they are all welfare recipients at randomization); and ii) the data and survey instruments gathered for all the individuals is the same. In the case of NEWWS the data available on each individual is extremely rich. By focusing in different geographic locations, however, we have the disadvantage of having to deal with the (potential) structural differences in local labor market, an issue that will be very important in this application.

Our strategy of comparing different control groups is similar to that previously used within a binary-treatment context by Friedlander and Robins (1995), Michalopoulos, Bloom, and Hill (2004) and Hotz, Imbens and Mortimer (2005). The key difference in our approach, however, is that while they focus on pairwise comparisons between controls in different locations, we focus on *simultaneously* comparing the individuals across *all* locations. This allows us to move beyond binary-treatment methods and evaluate nonexperimental estimators for multiple treatments because we need to adjust for differences in observed characteristics of several groups at the same time.

Finally, among the estimators we evaluate we pay particular attention to those based on the generalized propensity score or GPS (i.e., the probability of receiving a particular treatment conditional on covariates), such as weighting and partial-mean estimators. In addition, we systematically analyze the role of the GPS in identifying units across sites that are comparable in terms of observable characteristics, and provide guidance for its use in practice. We show the crucial role played by the GPS in extending to the multiple-treatment setting the "common support condition" frequently used in the binary-treatment setting.

The paper is organized as follows. Section 2 describes the data used in this paper. In section 3 we present the general set up, and in the following section we present the estimators to be used in the paper. The results are presented in Section 5, and Section 6 concludes.

2 Data

The data used in this paper comes from the National Evaluation of Welfare-to-Work Strategies (NEWWS), which is a multi-year study conducted in the early nineties to compare the effects of two alternative strategies to helping welfare recipients (mostly single mothers) to improve their labor market outcomes and leave public assistance. The first strategy 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. 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.[1] 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 program, for a pre-set "embargo" period.

The year in which random assignment took place differs across sites, with the earliest randomization starting in the second quarter of 1991 in Riverside, and the latest in the fourth quarter of 1994 in Portland.[2] The NEWWS data set contains information on labor market outcomes up to 5 years after random assignment, information on individual background characteristics, as well as individual welfare and labor market history up to two years prior to random assignment. We use these characteristics, further described in Section 5, to apply the nonexperimental estimators in which we will focus our analysis.[3]

As it will be explained in detail in the following section, we employ nonexperimental methods to eliminate differences in control group outcomes across the different locations in the NEWWS experiment. The total number of individuals in the control groups in the seven sites is 17,521. From these, we exclude all men from our analysis (1,303), and also all females with missing values on any of the variables used in the analysis (805). From the remaining observations, we also drop those controls for which it is unknown whether they were embargoed from the program services during the period considered (404). Finally, we exclude two sites, Columbus and Oklahoma City, from our analysis (5,658). Columbus has the problem of not having two years of labor market history prior to random assignment. Given the documented importance of controlling for such variables in nonexperimental settings (e.g., Heckman et al., 1997; Hotz et al., 2005) and the fact that it is the only site with that issue, we decided to exclude it from our analysis. We dropped Oklahoma City from the analysis because there randomization was done to welfare applicants, as opposed to welfare recipients as it was in the remaining sites. This implied that a big proportion (30%) of those

[1] One could use these sites to create alternative nonexperimental groups for those receiving LFA and HCD training. However, as discussed below, since LFA and HCD programs are heterogeneous across sites, this introduces additional biases. For this reason, we focus on comparing average outcomes for control individuals across sites, where everyone is excluded from receiving treatment. This also helps to increase the number of groups considered in our nonexperimental evaluation as the number of sites is greater than the number of alternative treatments.

[2] The dates in which randomization took place in all seven sites are (month/year): Atlanta (01/92-06/93), Columbus (09/92-07/94), Detroit (05/92-06/94), Grand Rapids (09/91-01/94), Oklahoma City (09/91-05/93), Portland (02/93-12/94) and Riverside (06/91-06/93).

[3] For further details on the NEWWS study see Hamilton et al. (2001).

individuals randomized actually did not qualify for welfare, and it is hard to believe they would be a good comparison group for individuals that did qualify. There is evidence in the literature that applicants and recipients are actually very different in terms of their characteristics and outcomes (e.g., Friedlander, 1988). Hence, in order to have groups across sites that are all formed by welfare recipients at randomization, we dropped Oklahoma City from the analysis. The final sample size in our analysis is 9,351 women, with 1,372 women from Atlanta; 3,037 from Detroit; 1,374 from Grand Rapids; 1,740 from Portland and 2,828 from Riverside.

The outcome we analyze in section 5 is the number of quarters employed during the two years following randomization and some variation on this outcome, explained below. We focus on an outcome measured two years after random assignment because in some sites we cannot be sure that all individuals were embargoed from receiving services from the program starting in year three.

3 General Framework

We base our general framework on the potential outcome approach developed by Neyman (1923) and extended by Rubin (1974) to non-experimental settings. Each unit i in our sample, $i = 1, 2, \dots, N$, comes from one of k possible sites. Let $D_i \in \{1, 2, \dots, k\}$ 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 location. Hence, $Y_i(t_d, d)$ is the outcome unit i would obtain if she were located in site d and given treatment t_d. Two differences with respect to the commonly used potential outcomes in program evaluation (e.g., Imbens, 2004) are worth mentioning. First, we let the potential outcome $Y(t_d, d)$ to depend on d for notational convenience. Although it may be difficult to think of the site 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 simultaneously use individuals from one site as a comparison group for another site. Second, we let t_d depend on d, as not all sites offered LFA and HCD training. For all sites, a value of t of zero denotes the control treatment, which prevents individuals from receiving any program services.

In this paper we focus exclusively on the control groups, so we use only the potential outcomes at zero, or $Y(0, d)$. The reason we focus only on controls is that not every site offered the two programs based on LFA and HCD, and programs differed across sites in terms of implementation, particular services offered, administration, etc. By focusing on the control treatment we try to minimize treatment heterogeneity across sites, and it allow us to use more sites as they all have a control group.[4]

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

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

The object in (1) gives the expected outcome under the control treatment in location d for someone randomly selected from our entire sample. In cases where d represents different levels of the treatment (and the zero is omitted from the potential outcome), (1) is the dose-response function.

Even though the treatment is randomly assigned within each site, and therefore $E[Y_i(0,d)|D_i = d]$ is identified from the data for every site, $E[Y_i(0,d)]$ is not identified without further assumptions. In general, it is not possible to use the controls from one location as a comparison group from another because the distribution of the characteristics in all k locations may differ. In order to evaluate nonexperimental methods that adjust for observable characteristics with multiple treatments, we impose the following unconfoundedness or selection-on-observables assumption.

Assumption 1 (Unconfounded site) The site an individual belongs to is unconfounded given pre-treatment variables X_i, or

$$D_i \perp \{Y_i(0,d)\}_{d\in\{1,2,\dots,k\}} | X_i \tag{2}$$

This assumption states that, conditional on a set of covariates, the site an individual belongs to is independent of her potential outcomes. Assumption 1 is similar to that in Hotz et al. (2005) in the binary treatment case.

In addition, we impose an overlap assumption that guarantees that in infinite samples we are able to compare units across all k sites for all values of X.

Assumption 2 (Simultaneous Overlap) For all x and all d

$$0 < \Pr(D_i = d | X = x) \tag{3}$$

[4] As in Hotz et al. (2005), if one is able to adjust for control group outcomes across sites, the comparison of adjusted outcomes for nominally equal treatments across sites (e.g., LFA programs in different locations) may be interpreted as the effect of program heterogeneity across sites.

By applying iterated expectations we can write $\beta_d = E[E[Y_i(0,d)|X = x]]$, which combined with assumptions 1 and 2 implies we can write β_d as a function of observed data as:

$$\beta_d = E[E[Y_i|D_i = d, X = x]] \tag{4}$$

The goal in this paper is to use the nonexperimental estimators described in the following section to adjust for observable characteristics in order to eliminate differences in average outcomes for controls among the different locations in the NEWWS. As mentioned before, the key in our approach, is that we want to compare all locations simultaneously, as opposed to the focus in the prior literature of making pairwise comparison between locations. Hence, the hypothesis we test in section 5 is that

$$\beta_1 = \beta_2 = \ldots = \beta_d \tag{5}$$

The equalities in (5) form the basis of our analysis as they imply that once we control for covariates and integrate over the appropriate distribution of those covariates, the individuals in any of the k locations can be used as a comparison group for all other locations. It is important to note that the outer expectation in (4) is over the distribution of the covariates over all the population (i.e., over all locations), and not over the distribution of the covariates for any given location. Hence, (5) does not imply that the average potential outcome for controls in each location is the same across locations –i.e., it does not imply that $E[Y_i(0,d)|D_i = d] = E[Y_i(0,d)|D_i = f]$ for $d \neq f$.

In addition, it is important to note that the overlap assumption rules out the use of local economic conditions as covariates. The reason is that with a fixed number of sites, the probability of finding another site with the same local economic conditions is zero, so the overlap assumption is violated. Therefore, assumptions 1 and 2 imply that controlling for pre-treatment variables is enough to make individuals comparable across site, without the use of local economic conditions. Because of this reason, Hotz et al. (2005) also call Assumption 1 the “no macro-effects” assumption. Since local economic conditions are likely to play an important role even after controlling for observed characteristics, in the analysis in section 5 we also present results controlling for them.

4 Non-experimental Estimators

In this section we discuss the different estimators of β_d we consider in this paper to eliminate differences in control outcomes across all sites. For comparison, we start with 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:

$$\widehat{\beta}_d^{raw} = \frac{\sum\limits_{i=1}^{N} Y_i 1(D_i = d)}{\sum\limits_{i=1}^{N} 1(D_i = d)} \tag{6}$$

This estimator would be an unbiased estimate of β_d if the individuals were randomly assigned across different locations. Since the characteristics of the individuals vary between locations, this estimator is a biased estimate of β_d. We use this estimator as a starting point, and we aim at reducing this bias by adjusting for differences in observable characteristics across locations under assumptions 1 and 2.

The result in (4) 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). The regression 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 (4) is a linear regression of the form:

$$E[Y_i | D_i, X_i] = \sum_{j=1}^{k} \alpha_j \cdot 1(D_i = j) + \delta' x_i \tag{7}$$

where δ is the coefficient vector for the covariates. Let the estimated coefficients in (7) be given by $\widehat{\alpha}_j$ and $\widehat{\delta}$. Then, the OLS-based estimator of β_d is given by:

$$\widehat{\beta}_d^{pmX} = \widehat{\alpha}_d + \widehat{\delta}' \left(N^{-1} \textstyle\sum_{i=1}^{N} x_i \right) \tag{8}$$

We also consider a more flexible model of (7) which contains polynomials of the continuous covariates and various interactions. We denote this estimator by $\widehat{\beta}_d^{pmXflex}$.

Recently, part of the program evaluation literature has focus on more flexible ways to control for covariates. The main issue in controlling for the covariates without imposing any structure in the model is that if the dimension of X is large, then non-parametric methods become intractable because of the so-called curse of dimensionality.

The same problem arises in the binary-treatment case. In a seminal paper, Rosenbaum and Rubin (1983) showed 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, defined as the probability of being in the treatment group conditional on X. This result implies that we only need to adjust for a scalar variable, as opposed to adjusting for all pretreatment variables. Since the propensity score is rarely known in practice, it is usually estimated using a logit model with interactions and high order terms in X, which can provide a relatively good approximation to the true model (e.g., Rosenbaum and Rubin, 1983; Dehejia and Wahba, 1985).[5]

Imbens (2000) and Lechner (2001) extended the results in Rosenbaum and Rubin (1983) to the multivalued and multiple treatment setting, and Hirano and Imbens (2004) further extended them to the continuous treatment case. The main 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) reduces the dimension to one, just as in the binary case.

Following Imbens (2000), define the generalized propensity score or GPS as the probability of receiving a particular treatment (in our case, belonging to a particular site) conditional on the covariates:

$$r(d, x) = \Pr(D = d | X = x) \tag{9}$$

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 unconfoundedness (Assumption 1) we can estimate the average potential outcomes by conditioning solely on the GPS. In particular, in our

[5]Note that, similar to the binary-treatment case, the problem of nonparametrically estimating the regression function of the outcome on the treatment and the covariates is translated to nonparametrically estimating the GPS. In practice, however, it may be preferable to impose restrictions (such as linearity) on the GPS rather than directly on the outcome.

context the result in Imbens (2000) can be written as:

$$\begin{aligned} (i)\ \gamma(d,r) &\equiv E[Y(0,d)\,|r(d,X)=r] = E[Y_i|D=d, r(D,X)=r] \qquad (10)\\ (ii)\ E[Y(0,d)] &= E[\gamma(d, r(t,X))] \end{aligned}$$

Therefore, the GPS can be used to estimate $\beta_d = E[Y(0,d)]$ by following the two steps in (10). First, one estimates the conditional expectation of Y as a function of D and $R = r(D,X)$ (i.e., the probability an individual gets the treatment she actually received). Second, to estimate β_d, we average the conditional expectation $\gamma(d,r)$ over $R^d = r(d,X)$. Hence, the averaging takes place over the values of the propensity score at the location corresponding to the parameter we want to estimate, in this case site d. As stressed in Imbens (2000), note that the second averaging is done over R^d, and not R. In addition, contrary to the binary-treatment case, in the multivalued or multiple treatment setting the conditional expectation $\gamma(d,r)$ does not have a causal interpretation.

The result in (10) suggests estimating β_d using a partial mean. However, contrary to the partial mean estimated using the covariates directly, 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. As before, the regressor that is fixed in the second step is the site.

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

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

Let the estimated coefficients from this regression be denoted by a hat on top of the coefficient. Next, we estimate β_d as:

$$\widehat{\beta}_d^{pmGPS} = E[Y(0)] = \frac{1}{N}\sum_{i=1}^{N}[\widehat{\alpha}_d \cdot 1(D_i=d) + \widehat{\delta}_d \cdot 1(D_i=d)\cdot R_i^d + \widehat{\eta}_j \cdot 1(D_i=d)\cdot (R_i^d)^2]$$

Alternatively, following Newey (1994) and more recently Flores (2007), we consider a more flexible specification in which the first step estimator of the regression function is based on a nonparametric kernel estimator. However, instead of employing the usual Nadaraya-Watson estimator, we use a local polynomial of order one. This estimator has the advantage that it does not have the boundary bias problem the former has. Since in our case the treatment is not continuous as in Flores (2007), 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. To formalize the estimator, let $K(u)$ be a kernel function such that $\int K(u)\,du = 1$; let h be a bandwidth satisfying $h \to 0$ and $Nh \to \infty$ as $N \to \infty$; and, let $K_h(u) = h^{-1}K(u/h)$. Then, the nonparametric estimator of $\gamma(d,r)$ in (10), $\widehat{\gamma}(d,r;h)$ is given by:[6]

$$\widehat{\gamma}(d,r;h) = \frac{1}{N}\sum_{i=1}^{N} \frac{\{\widehat{s}_2(r,h) - \widehat{s}_1(r,h)(R_i - r)\} K_h(R_i - r) \cdot Y_i \cdot 1(D_i = d)}{\widehat{s}_2(r,h)\,\widehat{s}_0(r,h) - \widehat{s}_1(r,h)^2} \tag{11}$$

where

$$\widehat{s}_v(r,h) = \frac{1}{N}\sum_{i=1}^{N} (R_i - r)^v K_h(R_i - r) \cdot 1(D_i = d)$$

Based on (11), our nonparametric partial mean estimator of β_d is given by:

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

In the next section, we implement this approach by using an Epanechnikov kernel and select the bandwidth using Silverman's rule: $h = 1.06 \min\{\widehat{\sigma}, I/1.34\} N^{-1/5}$, where $\widehat{\sigma}$ is the standard deviation of R_i and I is the interquartile range (e.g., Härdle et al., 2004).[7]

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. Similar to the binary treatment case, in a multiple or multivalued treatment case one can weight the observations receiving a given treatment level t by the probability of receiving the treatment they actually received conditional on X (i.e., R_i). More specifically, applying the results in Imbens (2000) to our context we can write β_d as a function of the observed data as

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

where as before, $R_i = r(D_i, X_i)$. Based on this result, a possible estimator of β_d

[6] See, for instance, Wand and Jones (1995).

[7] In the following section we analyze the sensitivity of our results to the choice of bandwidth by looking at the different estimates we obtain by varying it.

is its sample analogue given by replacing $E[\cdot]$ by the empirical average $N^{-1}\sum_{i=1}^{N}\cdot$. However, similar to the binary case discussed in Imbens (2004), this estimator has the undesirable property that its weights do not necessarily add to one. An alternative is to normalize the weights to add to one. Thus, the estimator we use in this case is given by

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

where *ipw* stands for inverse probability weight estimator. Similar to the binary-treatment case, note that $\widehat{\beta}_d^{ipw}$ for $d = 1, \ldots, k$ can also be calculated from the weighted linear regression

$$E[Y_i|D_i] = \sum_{j=1}^{k}\beta_j^{ipw}\cdot 1(D_i = j), \tag{12}$$

with weights equal to

$$w_i = \sqrt{\frac{1}{R_i}}$$

Following Imbens (2004), we also consider an inverse probability weight estimator that adds covariates to the weighted regression in (12).[8] Hence, we first estimate the weighted regression

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

with weights w_i. Next, we estimate β_d using the estimated coefficients of this weighted regression as:[9]

$$\widehat{\beta}_d^{ipwX} = \widehat{\alpha}_d + \widehat{\delta}'\left(N^{-1}\textstyle\sum_{i=1}^{N} x_i\right) \tag{13}$$

So far we have ignored two important issues in the implementation of the approaches based on GPS: estimation of the GPS and imposition of the overlap restriction. As in the binary-treatment case, the correct model underlying the GPS is unknown, and a nonparametric approach to its estimation becomes infeasible as the number of covariates grows. In this paper we follow an analogous approach to the binary-treatment

[8] For a discussion of this estimator in the binary-treatment case see, for instance, Imbens and Wooldridge (2008).

[9] In the binary-treatment case this second step is not needed since the weighted regression includes a treatment indicator (and a constant), and the focus is on estimating the treatment effect. Since here the parameter of interest is the average potential outcome, this second step is needed.

setting and estimate the GPS using a flexible multinomial logit that includes interactions and higher order terms of the pretreatment variables.

The overlap condition in Assumption 2 is stronger than that of the binary-treatment case, as it requires that we find comparable individuals across all sites for all values of X. In practice, when working with a binary treatment the usual approach is to drop units in the treatment or control group for which it is not possible to find a comparable individual in the other treatment arm, i.e., drop those individuals whose propensity score does not overlap with the propensity score of those in the other treatment arm. Hence, by doing this one redefines the parameter of interest to be conditional on the subpopulation with common overlap on the GPS.

The general idea of overlap in the multivalued case is similar to that for the binary case, but since now we want to compare different treatments simultaneously, we need to find comparable individuals across all treatment groups for all different treatments. Let the overlap region with respect to treatment (in our case location) d be given by the subsample

$$Overlap_d = \left\{ i : R_i^d \in \left[\max_{j=1,\ldots,k} \left\{ \min_{\{q: D_q = j\}} R_q^d \right\}, \min_{j=1,\ldots,k} \left\{ \max_{\{q: D_q = j\}} R_q^d \right\} \right] \right\} \tag{14}$$

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

$$Overlap = \bigcap_{d=1}^{k} Overlap_d \tag{15}$$

All the estimators based on the GPS are applied within the overlap region given by (15). By restricting our attention to units within the overlap region, we guarantee that we are able to find comparable units in all other locations. In order to analyze the importance of comparing "comparable" units in the multivalued or multiple case, we also implement the non-GPS-based estimators discussed in this section using the entire sample as well as only those units in the overlap region.

In the next section we also incorporate local economic conditions into our analysis. The variables we consider are several growth rates explained below. To control for these variables, we first regress the outcomes on the local economic conditions and then we apply the methods discussed in this section using the residual from this regression as the outcome.

5 Results

In this section we concentrate our analysis on the outcome "number of quarters employed in the first two years after randomization", and a "*differences*" version of it, in which we subtract the number of quarters employed in the two years prior to random assignment. As explained in Section 2, we concentrate on the control groups in five locations: Atlanta, Detroit, Grand Rapids, Portland and Riverside. Table 1 shows the descriptive statistics of the outcomes and covariates in each of these sites. 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. In addition, at the end of the table we present the variables that we use to control for local economic conditions, as explained above. These variables are the rates of growth (expressed in logs as the log of the variable one year minus the log of the same variable the previous year) of employment to population ratio, average real earnings and unemployment rate in the metropolitan statistical area (MSA) of each site. We use the information for years one and two after random assignment.

As expected, there are important and statistically significant differences across sites. For instance, while the percentage of blacks in Atlanta is 95 percent, this percentage is only 17 percent in Riverside. Also, individuals in Riverside appear to have better employment attachment and earnings histories, higher education level and less history of dependence on welfare and food stamps aid.

The second panel of Table 1 presents the same information, but after the overlap or common support condition in (15) is applied. The bottom of the two panels in the table show that 1,452 out of 9,351 units (over 15%) do not satisfy the conditions and therefore are dropped from all analyses where overlap is imposed. In general, it can be seen that for most variables the mean values of the covariates per site get closer to each other after imposing overlap, but in most cases not enough to eliminate the differences across sites.

As mentioned in Section 4, we estimated the GPS using a multinomial logit model. All individual level covariates presented in Table 1 were included in the estimation. We use this estimated GPS to further study how well covariates are "balanced" across sites. We follow two strategies. The first one, for which the results are presented in panel A of Table 2, tests for *each covariate* if there is joint equality of means across all sites. Clearly, imposing overlap by itself does not make a difference, for literally in all covariates the tests are rejected. However, when we perform the same test weighting each observation by the inverse of the GPS (inverse probability weighting), only 6 out of 53 covariates appear as not balanced at the 5% significance level. This suggests that the GPS is (mostly) attaining the desired result.

The second approach, presented in panel B of Table 2, consists of a series of pairwise comparisons of the means of each site versus all other sites. The two "raw means" versions (before and after imposing the overlap condition) just test equality of means. The third version consists of dividing ("blocking") the units in each site by the decile of the GPS in their site and calculating the difference of means with all the units in other sites for which their estimated GPS for that particular site is in the same decile. For example, for individuals in Atlanta in the first decile of the estimated GPS for individuals living in Atlanta, we chose as comparison group all the individuals living in other sites for which their GPS of being in Atlanta is in the same first decile. The weighted average (by the number of individuals) of these difference of means (and the corresponding standard error) are used to test the equality of means of each site versus the other sites. Here, Table 2 shows that the results are mixed. On one hand, for most sites the number of covariates with significant differences decreases with the application of blocking. On the other hand, some issues remain, for example, with Detroit.

In summary, it appears that in general the GPS is helping in attaining balancing of covariates, at least for a large number of them and for most sites. In Appendix Table 1 we present detailed information on the tests used to generate Table 2. For the blocking estimator, we present the standardized (by the average and the standard deviation) difference of means, and indicate the significance level of those differences.

Next we calculate all the estimators presented in Section 4 on four outcomes. First, we use the number of quarters employed in the two years after random assignment as outcome. Second, we use the differences outcome described above. The results for these two outcomes are presented in Table 3 and in Figures 1a, 1b, 2a and 2b. The final two outcomes considered are the residuals from the regression where we try to eliminate the portion of the outcomes explained by just differences in local economic conditions. The regressions only include a constant and the six local economic conditions variables presented at the bottom of Table 1. We refer to these residuals with the same name as the original outcome, but with the suffix "adjusted by LEC". The results for these adjusted outcomes are presented in Table 4 and in figures 3a, 3b, 4a and 4b.

In Tables 3 and 4 and all the figures, the confidence intervals were estimated by bootstrapping with 500 replications. Also, in the tables and graphs we present the p-values of the joint test that the estimated expected value of the outcomes across sites are equal.

We can observe from the figures and Tables 3 and 4 that for the outcome in levels all the estimators have trouble trying to make the expected values equal across sites. The partial mean estimators, however, work reasonably well in adjusting outcomes in most of the sites. For the differences outcome the results are similar (see Figures 2a and 2b), but clearly just the fact of taking the difference improves the similarity of the mean outcomes across sites.

Finally, when the outcomes are adjusted first by local economic conditions, then equality of means is attained for many of the estimators, including all the GPS-based ones. This implies that once the structural differences in the labor markets for each site are accounted for, then the estimators work very well in eliminating any remaining differences due to individual-level factors.

6 Conclusion

[To be completed]

References

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

[2] Cattaneo, Matias. 2007. Efficient Semiparametric Estimation of Multi-valued Treatment Effects. University of California, Berkeley, November.

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

[4] 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.

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

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

[7] Flores-Lagunes, Alfonso, Arturo Gonzalez, and Todd C. Neumann. 2007. Estimating the Effects of Length of Exposure to a Training Program: The Case of Job Corps. IZA Discussion Paper, no. 2846 (June).

[8] Friedlander, Daniel (1988). Subgroups Impacts and Performance Indicators for Selected Welfare Employment Programs. New York: Manpower Demonstration Research Corporation.

[9] 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.

[10] 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: 375-396.

[11] Hamilton, G.; Freedman, S.; Gennetian, L.; Michalopoulos, C.; Walter, J.; Adams-Ciardullo, D.; Gassman-Pines, A.; McGroder, S.; Zaslow, M. Brooks. J. and Ahluwalia S. (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.

[12] Härdle, W.; Müller, M.; Sperlich, S. and Werwatz, A. (2004). Nonparametric and Semiparametric Models. Springer Series in Statistics.

[13] 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.

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

[15] Heckman, James J., Hidehiko Ichimura, J. Smith and Petra E. Todd. 1998b. Characterizing Selection Bias Using Experimental Data. Econometrica, 66, 1017-1098.

[16] Heckman, J. and J. Hotz (1989). Alternative Methods for Evaluating the Impact of Training Programs, (with discussion), Journal of the American Statistical Association, 84, 804, 862-874.

[17] Heckman, J., LaLonde, R. and Smith, J. (1999) "The Economics and Econometrics of Active Labor Market Programs" in O. Ashenfelter and D. Card (eds.) Handbook of Labor Economics. Elsevier Science North Holland, 1865-2097.

[18] Hirano, K. and Imbens, G. W. (2004). “The Propensity Score with Continuous Treatments”, in Andrew Gelman and Xiao-Li Meng eds., Applied Bayesian

Modeling and Causal Inference from Incomplete-Data Perspectives. John Wiley and Sons, 73-84.

[19] Holland, P. W. (1986). "Statistics and Causal Inference"(with Discussion and Reply), Journal of the American Statistical Association, 81(396), 945-970.

[20] 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: 241-270.

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

[22] Imbens, G. (2004) "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review" Review of Economics and Statistics, 84, 4-29.

[23] Imbens G. and Wooldridge, J. (2008). Recent Developments in the Econometrics of Program Evaluation, National Bureau of Economic Research Working Paper #14251.

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

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

[26] 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.

[27] 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? Review of Economics and Statistics 86, no. 1: 156-179.

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

[29] Newey, W. K. (1994). "Kernel Estimation of Partial Means and a General Variance Estimator", Econometric Theory, 10, 233-253.

[30] Neyman, J. (1923) "On the Application of Probability Theory to Agricultural Experiments: Essays on Principles" Translated in Statistical Science, 5, 465-80.

[31] Rubin, D. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies" Journal of Educational Psychology, 66, 688-701.

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

[33] Wand, M. P. and Jones, M. C. (1995). Kernel Smoothing, Chapman and Hall, London, New York.

Table 1. Descriptive Statistics NEWWS Data

Variables	Before Imposing Overlap Conditon					After Imposing Overlap Conditon				
	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV
Outcomes										
# qtrs employed in 2 years after RA	2.48	2.21	2.80	2.51	1.97	2.47	2.24	2.71	2.46	1.88
	(2.82)	(2.49)	(2.62)	(2.78)	(2.71)	(2.81)	(2.49)	(2.59)	(2.76)	(2.64)
# qtrs employed in 2 years after RA (Diff)	0.62	0.78	0.23	0.52	-0.17	0.69	0.80	0.31	0.56	-0.10
	(2.64)	(2.60)	(2.87)	(2.90)	(2.91)	(2.60)	(2.60)	(2.79)	(2.83)	(2.85)
Covariates										
Demographic & Family Characteristics										
Black	0.95	0.89	0.41	0.20	0.17	0.95	0.89	0.46	0.23	0.24
	(0.22)	(0.32)	(0.49)	(0.40)	(0.38)	(0.22)	(0.32)	(0.50)	(0.42)	(0.43)
Age 30-39 years old	0.51	0.35	0.29	0.40	0.45	0.50	0.35	0.31	0.40	0.44
	(0.50)	(0.48)	(0.46)	(0.49)	(0.50)	(0.50)	(0.48)	(0.46)	(0.49)	(0.50)
Age 40+ years old	0.14	0.11	0.09	0.08	0.13	0.13	0.11	0.09	0.08	0.13
	(0.34)	(0.32)	(0.28)	(0.27)	(0.34)	(0.34)	(0.31)	(0.29)	(0.28)	(0.34)
Teenage mother (at <=19 years)	0.45	0.45	0.51	0.34	0.35	0.46	0.46	0.51	0.35	0.37
	(0.50)	(0.50)	(0.50)	(0.47)	(0.48)	(0.50)	(0.50)	(0.50)	(0.48)	(0.48)
Never married	0.62	0.69	0.58	0.49	0.34	0.63	0.69	0.59	0.52	0.38
	(0.48)	(0.46)	(0.49)	(0.50)	(0.47)	(0.48)	(0.46)	(0.49)	(0.50)	(0.49)
Any child 0-5 years old	0.42	0.65	0.69	0.71	0.58	0.44	0.66	0.67	0.69	0.59
	(0.49)	(0.48)	(0.46)	(0.46)	(0.49)	(0.50)	(0.47)	(0.47)	(0.46)	(0.49)
Any child 6-12 years old	0.70	0.48	0.43	0.52	0.59	0.69	0.48	0.45	0.53	0.58
	(0.46)	(0.50)	(0.49)	(0.50)	(0.49)	(0.46)	(0.50)	(0.50)	(0.50)	(0.49)
Two children in household	0.34	0.30	0.36	0.33	0.32	0.34	0.30	0.36	0.34	0.33
	(0.47)	(0.46)	(0.48)	(0.47)	(0.47)	(0.47)	(0.46)	(0.48)	(0.47)	(0.47)
Three or more children in household	0.31	0.27	0.19	0.30	0.28	0.31	0.27	0.20	0.29	0.28
	(0.46)	(0.44)	(0.39)	(0.46)	(0.45)	(0.46)	(0.45)	(0.40)	(0.45)	(0.45)
Education Characteristics										
10th grade	0.14	0.15	0.13	0.17	0.11	0.13	0.14	0.13	0.17	0.12
	(0.35)	(0.35)	(0.34)	(0.38)	(0.31)	(0.34)	(0.35)	(0.34)	(0.38)	(0.33)
11th grade	0.17	0.25	0.20	0.22	0.18	0.18	0.26	0.21	0.21	0.17
	(0.38)	(0.44)	(0.40)	(0.41)	(0.38)	(0.38)	(0.44)	(0.40)	(0.40)	(0.38)
Grade 12 or higher	0.57	0.50	0.54	0.45	0.57	0.58	0.50	0.53	0.45	0.55
	(0.49)	(0.50)	(0.50)	(0.50)	(0.49)	(0.49)	(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	0.52	0.55
	(0.50)	(0.50)	(0.50)	(0.50)	(0.49)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)

(continues in next page)

Table 1. Descriptive Statistics NEWWS Data (continuation)

Variables	Before Imposing Overlap Conditon					After Imposing Overlap Conditon				
	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV
Housing Type & Housing Stability										
Lives in public/subsidized house	0.59	0.07	0.16	0.29	0.09	0.57	0.07	0.17	0.28	0.10
	(0.49)	(0.26)	(0.37)	(0.46)	(0.29)	(0.49)	(0.26)	(0.38)	(0.45)	(0.30)
One or two moves in past 2 years	0.49	0.48	0.51	0.47	0.54	0.49	0.49	0.53	0.48	0.53
	(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	0.08	0.22	0.20	0.21
	(0.27)	(0.27)	(0.43)	(0.42)	(0.41)	(0.27)	(0.27)	(0.41)	(0.40)	(0.41)
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	0.31	0.40
	(0.44)	(0.42)	(0.49)	(0.47)	(0.50)	(0.43)	(0.42)	(0.48)	(0.46)	(0.49)
On welfare for 2-5 years	0.25	0.25	0.31	0.35	0.28	0.25	0.26	0.32	0.36	0.29
	(0.43)	(0.44)	(0.46)	(0.48)	(0.45)	(0.43)	(0.44)	(0.47)	(0.48)	(0.46)
On welfare 5-10 years	0.24	0.24	0.17	0.23	0.16	0.24	0.24	0.18	0.23	0.17
	(0.43)	(0.43)	(0.38)	(0.42)	(0.37)	(0.43)	(0.43)	(0.39)	(0.42)	(0.37)
Received Welfare in Q1 before RA	0.97	0.90	0.77	0.79	0.73	0.97	0.91	0.83	0.85	0.79
	(0.18)	(0.29)	(0.42)	(0.41)	(0.44)	(0.18)	(0.29)	(0.37)	(0.36)	(0.41)
Received Welfare in Q2 before RA	0.93	0.86	0.70	0.74	0.49	0.93	0.86	0.76	0.80	0.61
	(0.26)	(0.35)	(0.46)	(0.44)	(0.50)	(0.26)	(0.35)	(0.43)	(0.40)	(0.49)
Received Welfare in Q3 before RA	0.85	0.84	0.68	0.72	0.46	0.85	0.84	0.73	0.77	0.57
	(0.36)	(0.37)	(0.47)	(0.45)	(0.50)	(0.35)	(0.37)	(0.44)	(0.42)	(0.49)
Received Welfare in Q4 before RA	0.73	0.83	0.67	0.69	0.44	0.77	0.83	0.71	0.74	0.52
	(0.44)	(0.38)	(0.47)	(0.46)	(0.50)	(0.42)	(0.38)	(0.45)	(0.44)	(0.50)
Received Welfare in Q5 before RA	0.69	0.81	0.64	0.64	0.41	0.72	0.81	0.68	0.68	0.49
	(0.46)	(0.40)	(0.48)	(0.48)	(0.49)	(0.45)	(0.40)	(0.47)	(0.47)	(0.50)
Received Welfare in Q6 before RA	0.66	0.79	0.61	0.61	0.39	0.69	0.79	0.64	0.65	0.46
	(0.47)	(0.41)	(0.49)	(0.49)	(0.49)	(0.46)	(0.41)	(0.48)	(0.48)	(0.50)
Received Welfare in Q7 before RA	0.64	0.77	0.56	0.58	0.37	0.67	0.77	0.60	0.61	0.43
	(0.48)	(0.42)	(0.50)	(0.49)	(0.48)	(0.47)	(0.42)	(0.49)	(0.49)	(0.50)
Food Stamps Use History										
Received FS in Q1 before RA	0.97	0.94	0.85	0.86	0.62	0.97	0.94	0.90	0.91	0.75
	(0.17)	(0.23)	(0.36)	(0.35)	(0.48)	(0.17)	(0.23)	(0.30)	(0.28)	(0.43)
Received FS in Q2 before RA	0.95	0.89	0.76	0.81	0.42	0.95	0.89	0.83	0.88	0.59
	(0.22)	(0.31)	(0.43)	(0.39)	(0.49)	(0.23)	(0.31)	(0.38)	(0.33)	(0.49)
Received FS in Q3 before RA	0.90	0.87	0.72	0.79	0.39	0.90	0.87	0.77	0.84	0.54
	(0.30)	(0.34)	(0.45)	(0.41)	(0.49)	(0.30)	(0.34)	(0.42)	(0.36)	(0.50)

(continues in next page)

Table 1. Descriptive Statistics NEWWS Data (continuation)

Variables	Before Imposing Overlap Conditon					After Imposing Overlap Conditon				
	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV
Food Stamps Use History (continued)										
Received FS in Q4 before RA	0.83	0.86	0.72	0.76	0.36	0.83	0.86	0.76	0.82	0.49
	(0.38)	(0.35)	(0.45)	(0.43)	(0.48)	(0.37)	(0.35)	(0.43)	(0.39)	(0.50)
Received FS in Q5 before RA	0.78	0.83	0.67	0.72	0.33	0.79	0.83	0.71	0.77	0.45
	(0.42)	(0.38)	(0.47)	(0.45)	(0.47)	(0.41)	(0.38)	(0.45)	(0.42)	(0.50)
Received FS in Q6 before RA	0.75	0.81	0.64	0.70	0.31	0.76	0.80	0.68	0.74	0.42
	(0.43)	(0.40)	(0.48)	(0.46)	(0.46)	(0.43)	(0.40)	(0.47)	(0.44)	(0.49)
Received FS in Q7 before RA	0.72	0.78	0.60	0.66	0.29	0.73	0.78	0.63	0.70	0.39
	(0.45)	(0.41)	(0.49)	(0.47)	(0.45)	(0.45)	(0.41)	(0.48)	(0.46)	(0.49)
Year of Random Assignment										
Random Assignment in year 1993	0.34	0.50	0.36	0.74	0.26	0.33	0.49	0.36	0.75	0.32
	(0.47)	(0.50)	(0.48)	(0.44)	(0.44)	(0.47)	(0.50)	(0.48)	(0.43)	(0.47)
Employment History										
Employed in Q1 before RA	0.18	0.18	0.29	0.23	0.22	0.18	0.18	0.27	0.21	0.19
	(0.39)	(0.38)	(0.45)	(0.42)	(0.41)	(0.38)	(0.39)	(0.45)	(0.41)	(0.39)
Employed in Q2 before RA	0.18	0.18	0.30	0.25	0.25	0.18	0.18	0.28	0.23	0.22
	(0.38)	(0.38)	(0.46)	(0.43)	(0.43)	(0.38)	(0.39)	(0.45)	(0.42)	(0.41)
Employed in Q3 before RA	0.19	0.18	0.29	0.25	0.26	0.18	0.18	0.27	0.23	0.23
	(0.39)	(0.38)	(0.46)	(0.43)	(0.44)	(0.39)	(0.38)	(0.44)	(0.42)	(0.42)
Employed in Q4 before RA	0.22	0.17	0.30	0.24	0.28	0.21	0.18	0.27	0.22	0.25
	(0.41)	(0.38)	(0.46)	(0.42)	(0.45)	(0.41)	(0.38)	(0.45)	(0.42)	(0.43)
Employed in Q5 before RA	0.24	0.17	0.31	0.24	0.28	0.23	0.18	0.29	0.23	0.26
	(0.43)	(0.38)	(0.46)	(0.43)	(0.45)	(0.42)	(0.38)	(0.45)	(0.42)	(0.44)
Employed in Q6 before RA	0.27	0.18	0.34	0.25	0.29	0.25	0.18	0.31	0.24	0.27
	(0.44)	(0.38)	(0.47)	(0.43)	(0.45)	(0.43)	(0.39)	(0.46)	(0.43)	(0.44)
Employed in Q7 before RA	0.29	0.18	0.36	0.26	0.29	0.27	0.18	0.34	0.26	0.28
	(0.45)	(0.39)	(0.48)	(0.44)	(0.45)	(0.45)	(0.39)	(0.48)	(0.44)	(0.45)
Employed in Q8 before RA	0.30	0.18	0.39	0.27	0.30	0.28	0.18	0.36	0.27	0.29
	(0.46)	(0.38)	(0.49)	(0.45)	(0.46)	(0.45)	(0.38)	(0.48)	(0.44)	(0.45)
Employed at RA (self reported)	0.07	0.07	0.13	0.09	0.13	0.07	0.07	0.13	0.08	0.11
	(0.26)	(0.25)	(0.34)	(0.28)	(0.33)	(0.26)	(0.26)	(0.33)	(0.27)	(0.31)
Ever worked FT 6+ months at same job	0.72	0.46	0.64	0.77	0.71	0.71	0.47	0.63	0.75	0.69
	(0.45)	(0.50)	(0.48)	(0.42)	(0.45)	(0.45)	(0.50)	(0.48)	(0.43)	(0.46)

(continues in next page)

Table 1. Descriptive Statistics NEWWS Data (continuation)

Variables	Before Imposing Overlap Conditon					After Imposing Overlap Conditon				
	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV
Earnings History (real $ /1,000)										
Earnings Q1 before RA	0.23	0.21	0.36	0.33	0.43	0.21	0.21	0.31	0.28	0.33
	(0.82)	(0.68)	(1.06)	(0.89)	(1.23)	(0.72)	(0.68)	(0.98)	(0.79)	(1.00)
Earnings Q2 before RA	0.26	0.25	0.52	0.41	0.63	0.25	0.25	0.45	0.34	0.46
	(0.85)	(0.82)	(1.29)	(1.04)	(1.55)	(0.77)	(0.83)	(1.20)	(0.93)	(1.19)
Earnings Q3 before RA	0.29	0.26	0.55	0.41	0.72	0.26	0.26	0.46	0.35	0.55
	(0.92)	(0.89)	(1.33)	(1.07)	(1.73)	(0.81)	(0.90)	(1.22)	(0.98)	(1.43)
Earnings Q4 before RA	0.41	0.25	0.53	0.43	0.74	0.37	0.25	0.47	0.38	0.58
	(1.22)	(0.82)	(1.29)	(1.14)	(1.68)	(1.12)	(0.83)	(1.23)	(1.06)	(1.40)
Earnings Q5 before RA	0.51	0.29	0.57	0.46	0.79	0.45	0.29	0.52	0.41	0.64
	(1.27)	(0.94)	(1.32)	(1.16)	(1.82)	(1.13)	(0.94)	(1.28)	(1.09)	(1.55)
Earnings Q6 before RA	0.62	0.31	0.62	0.51	0.80	0.55	0.31	0.57	0.46	0.69
	(1.44)	(1.01)	(1.41)	(1.26)	(1.81)	(1.32)	(1.01)	(1.39)	(1.20)	(1.60)
Earnings Q7 before RA	0.72	0.32	0.68	0.54	0.83	0.64	0.33	0.64	0.51	0.73
	(1.65)	(1.06)	(1.44)	(1.31)	(1.89)	(1.53)	(1.07)	(1.42)	(1.30)	(1.66)
Earnings Q8 before RA	0.74	0.33	0.69	0.57	0.85	0.66	0.34	0.65	0.55	0.76
	(1.61)	(1.09)	(1.45)	(1.35)	(1.86)	(1.46)	(1.10)	(1.45)	(1.32)	(1.73)
Any earnings year before RA (self-rep)	0.23	0.20	0.46	0.36	0.40	0.23	0.21	0.42	0.33	0.36
	(0.42)	(0.40)	(0.50)	(0.48)	(0.49)	(0.42)	(0.40)	(0.49)	(0.47)	(0.48)
Local Economic Conditions (growth rates)										
Employment/Population year 1 after RA	0.02	0.02	0.02	0.02	-0.01	0.02	0.02	0.02	0.02	-0.01
	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
Average Total earnings year 1 after RA	-0.01	0.02	0.01	0.01	-0.01	-0.01	0.02	0.01	0.01	-0.01
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Unemployment Rate year 1 after RA	-0.15	-0.21	-0.19	-0.23	0.01	-0.15	-0.21	-0.19	-0.23	0.00
	(0.01)	(0.05)	(0.10)	(0.05)	(0.10)	(0.01)	(0.05)	(0.10)	(0.05)	(0.10)
Employment/Population year 2 after RA	0.02	0.02	0.03	0.02	0.01	0.02	0.02	0.03	0.02	0.01
	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.01)
Average Total earnings year 2 after RA	0.00	0.01	0.01	0.02	-0.01	0.00	0.01	0.01	0.02	-0.01
	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)
Unemployment Rate year 2 after RA	-0.13	-0.12	-0.21	-0.08	-0.10	-0.13	-0.12	-0.21	-0.08	-0.10
	(0.04)	(0.07)	(0.02)	(0.12)	(0.06)	(0.04)	(0.07)	(0.02)	(0.12)	(0.06)
Number of observations per site	1,372	2,037	1,374	1,740	2,828	1,290	1,976	1,189	1,496	1,948
Total number of observations			9,351					7,899		

Table 2. Summary Results from Balancing of Covariates Analysis

A. Joint tests of equality of means of covariates across all sites

Method	Number of covariates for which		
	P-Value ≤ 0.10	P-Value ≤ 0.05	P-Value ≤ 0.01
Raw Means Before Overlap	53	53	53
Raw Means After Overlap	53	53	52
GPS-based Inverse Probability Weighting	10	6	0
Total Number of Covariates		53	

B. Tests of differences of means of covariates in one site vs all other sites pooled together

Method	Number of covariates for which		
	P-Value ≤ 0.10	P-Value ≤ 0.05	P-Value ≤ 0.01
Raw Means Before Overlap			
Atlanta vs others	43	43	40
Detroit vs others	50	50	49
Grand Rapids vs others	38	35	31
Portland vs others	38	37	30
Riverside vs others	50	49	46
Raw Means After Overlap			
Atlanta vs others	41	41	36
Detroit vs others	49	41	36
Grand Rapids vs others	36	30	24
Portland vs others	31	26	20
Riverside vs others	48	48	44
Blocking on GPS			
Atlanta vs others	18	16	12
Detroit vs others	35	30	14
Grand Rapids vs others	16	9	5
Portland vs others	16	12	7
Riverside vs others	16	13	6
Total Number of Covariates		53	

Table 3. Estimated Average Outcomes per Site and Bootstrap Confidence Intervals
Outcome: Number of Quarters Employed in Two Years after Random Assignment

Estimator	Outcome in Levels					Joint Equality	Outcome in Differences					Joint Equality
	ATL	DET	GRP	POR	RIV	Test (p-value)	ATL	DET	GRP	POR	RIV	Test (p-value)
RAW_NO_OV	2.5	2.2	2.8	2.5	2.0	0.000	0.6	0.8	0.2	0.5	-0.2	0.000
	[2.4,2.5]	[2.2,2.3]	[2.7,2.8]	[2.5,2.6]	[1.9,2.0]		[0.6,0.7]	[0.8,0.8]	[0.2,0.3]	[0.5,0.6]	[-0.3,-0.1]	
RAW_OV	2.6	2.3	2.6	2.4	1.8	0.000	0.8	0.8	0.5	0.8	0.1	0.000
	[2.5,2.6]	[2.3,2.3]	[2.6,2.7]	[2.4,2.6]	[1.7,1.8]		[0.7,0.8]	[0.8,0.9]	[0.5,0.5]	[0.8,0.9]	[0.0,0.3]	
Covariates-Based												
PM_X_NO_OV	2.5	2.3	2.5	2.5	1.8	0.000	0.8	0.6	0.7	0.7	0.1	0.000
	[2.4,2.5]	[2.3,2.4]	[2.4,2.5]	[2.5,2.6]	[1.6,1.8]		[0.8,0.8]	[0.6,0.7]	[0.6,0.7]	[0.7,0.7]	[-0.1,0.1]	
PM_X_OV	4.0	3.8	3.8	4.0	3.2	0.000	2.2	1.9	1.9	2.0	1.4	0.000
	[3.8,4.4]	[3.6,4.2]	[3.6,4.3]	[3.8,4.4]	[3.0,3.5]		[2.2,2.4]	[2.0,2.1]	[1.9,2.1]	[2.1,2.3]	[1.4,1.5]	
PM_X_FLEX_NO_OV	3.7	3.5	3.6	3.7	3.0	0.000	1.7	1.6	1.6	1.7	1.0	0.000
	[3.5,4.0]	[3.3,3.9]	[3.3,3.9]	[3.5,4.0]	[2.7,3.1]		[1.5,2.0]	[1.4,1.9]	[1.4,1.9]	[1.5,2.0]	[0.8,1.1]	
PM_X_FLEX_OV	4.0	3.7	3.7	3.9	3.2	0.000	2.1	1.8	1.8	2.0	1.3	0.000
	[3.9,4.4]	[3.6,4.2]	[3.6,4.1]	[3.7,4.2]	[3.0,3.5]		[2.1,2.4]	[1.8,2.2]	[1.8,2.2]	[1.9,2.3]	[1.2,1.5]	
GPS-Based												
PM_GPS_PAR_OV	2.6	2.3	2.3	2.6	1.9	0.001	0.8	0.6	0.6	0.9	0.2	0.351
	[2.6,2.7]	[2.2,2.3]	[2.3,2.4]	[2.5,2.9]	[1.7,1.8]		[0.8,0.9]	[0.6,0.7]	[0.6,0.6]	[0.7,1.1]	[0.1,0.3]	
PM_GPS_NPR_OV	2.3	2.6	2.2	2.5	1.9	0.805	0.7	0.6	0.5	1.0	0.3	0.256
	[2.6,2.8]	[2.3,2.6]	[2.3,2.4]	[2.2,2.7]	[1.5,3.4]		[0.6,0.8]	[0.4,0.9]	[0.4,0.5]	[0.6,1.2]	[0.1,0.5]	
IPW_OV	2.5	2.4	2.2	2.4	1.8	0.774	0.7	0.7	0.6	0.9	0.1	0.499
	[2.4,2.6]	[2.2,2.4]	[2.3,2.4]	[2.3,2.8]	[1.7,1.9]		[0.7,0.7]	[0.5,0.7]	[0.5,0.6]	[0.7,1.2]	[0.0,0.1]	
IPW_X_OV	2.5	2.1	2.8	2.5	1.9	0.000	0.5	0.7	0.4	0.8	0.1	0.000
	[2.5,2.6]	[2.0,2.2]	[2.8,2.8]	[2.4,2.7]	[1.7,1.9]		[0.5,0.7]	[0.6,0.8]	[0.4,0.5]	[0.7,1.0]	[-0.2,0.1]	

Notes: Bootstrap Confidence Intervals between brackets (based on 500 replications).
The outcome in differences substracts the outcome in years 1 and 2 before Random Assignment from the outcome in levels

Table 4. Estimated Average Outcomes per Site and Bootstrap Confidence Interval
Outcome: Number of Quarters Employed in Two Years after Random Assignment Adjusted by Local Economic Conditions After RA

Estimator	Outcome in Levels					Joint Equality	Outcome in Differences					Joint Equality
	ATL	DET	GRP	POR	RIV	Test (p-value)	ATL	DET	GRP	POR	RIV	Test (p-value)
RAW_NO_OV	0.1	-0.2	0.2	0.0	-0.1	0.008	0.1	0.2	-0.1	-0.1	-0.1	0.186
	[0.1,0.2]	[-0.1,-0.1]	[0.1,0.3]	[-0.1,0.0]	[-0.2,-0.1]		[0.0,0.1]	[0.2,0.3]	[-0.1,0.0]	[-0.1,0.0]	[-0.2,0.0]	
RAW_OV	0.2	-0.1	0.0	-0.1	-0.2	0.000	0.2	0.2	0.2	0.1	0.2	0.992
	[0.2,0.3]	[-0.1,0.0]	[0.0,0.1]	[-0.2,0.0]	[-0.3,-0.2]		[0.2,0.3]	[0.2,0.3]	[0.2,0.2]	[0.2,0.3]	[0.1,0.4]	
Covariates-Based												
PM_X_NO_OV	0.1	0.0	-0.1	0.0	-0.3	0.000	0.2	0.1	0.3	0.2	0.1	0.109
	[0.1,0.1]	[0.0,0.0]	[-0.2,-0.1]	[0.0,0.1]	[-0.4,-0.2]		[0.2,0.3]	[0.1,0.1]	[0.3,0.3]	[0.2,0.3]	[0.0,0.2]	
PM_X_OV	-0.6	-0.9	-1.1	-0.9	-1.2	0.000	1.4	1.1	1.3	1.2	1.2	0.000
	[-0.9,-0.2]	[-1.1,-0.4]	[-1.3,-0.6]	[-1.1,-0.5]	[-1.4,-0.8]		[1.4,1.5]	[1.2,1.3]	[1.3,1.5]	[1.3,1.5]	[1.2,1.3]	
PM_X_FLEX_NO_OV	-1.0	-1.2	-1.3	-1.2	-1.4	0.000	0.9	0.7	1.0	0.9	0.8	0.068
	[-1.2,-0.7]	[-1.4,-0.8]	[-1.6,-0.9]	[-1.4,-0.9]	[-1.6,-1.2]		[0.7,1.2]	[0.6,1.1]	[0.8,1.3]	[0.7,1.2]	[0.6,0.9]	
PM_X_FLEX_OV	-0.7	-1.0	-1.1	-1.0	-1.2	0.000	1.3	1.0	1.2	1.2	1.1	0.001
	[-0.8,-0.3]	[-1.1,-0.5]	[-1.3,-0.7]	[-1.2,-0.6]	[-1.4,-0.9]		[1.2,1.6]	[1.0,1.4]	[1.2,1.6]	[1.1,1.5]	[1.0,1.3]	
GPS-Based												
PM_GPS_PAR_OV	0.3	-0.1	-0.2	-0.1	-0.1	0.000	0.3	0.1	0.3	0.3	0.3	0.200
	[0.3,0.4]	[-0.2,0.0]	[-0.3,-0.2]	[-0.2,0.2]	[-0.4,-0.2]		[0.2,0.3]	[0.1,0.2]	[0.2,0.3]	[0.0,0.4]	[0.2,0.3]	
PM_GPS_NPR_OV	0.0	0.2	-0.4	-0.2	-0.2	0.772	0.2	0.0	0.2	0.3	0.4	0.757
	[0.2,0.5]	[-0.1,0.3]	[-0.3,-0.2]	[-0.4,0.1]	[-0.6,1.4]		[0.1,0.2]	[-0.1,0.3]	[0.0,0.2]	[0.0,0.6]	[0.1,0.6]	
IPW_OV	0.1	0.0	-0.4	-0.2	-0.2	0.001	0.1	0.1	0.2	0.2	0.1	0.290
	[0.1,0.3]	[-0.1,0.1]	[-0.3,-0.2]	[-0.4,0.2]	[-0.4,-0.2]		[0.1,0.2]	[0.0,0.2]	[0.2,0.2]	[0.0,0.5]	[0.1,0.2]	
IPW_X_OV	0.2	-0.2	0.2	-0.2	-0.1	0.000	0.0	0.2	0.1	0.0	0.2	0.199
	[0.2,0.3]	[-0.4,-0.1]	[0.2,0.2]	[-0.3,0.0]	[-0.4,-0.2]		[0.0,0.2]	[0.0,0.3]	[0.0,0.1]	[0.0,0.2]	[-0.1,0.2]	

Notes: Bootstrap Confidence Intervals between brackets (based on 500 replications).
The outcome in differences substracts the outcome in years 1 and 2 before Random Assignment from the outcome in levels

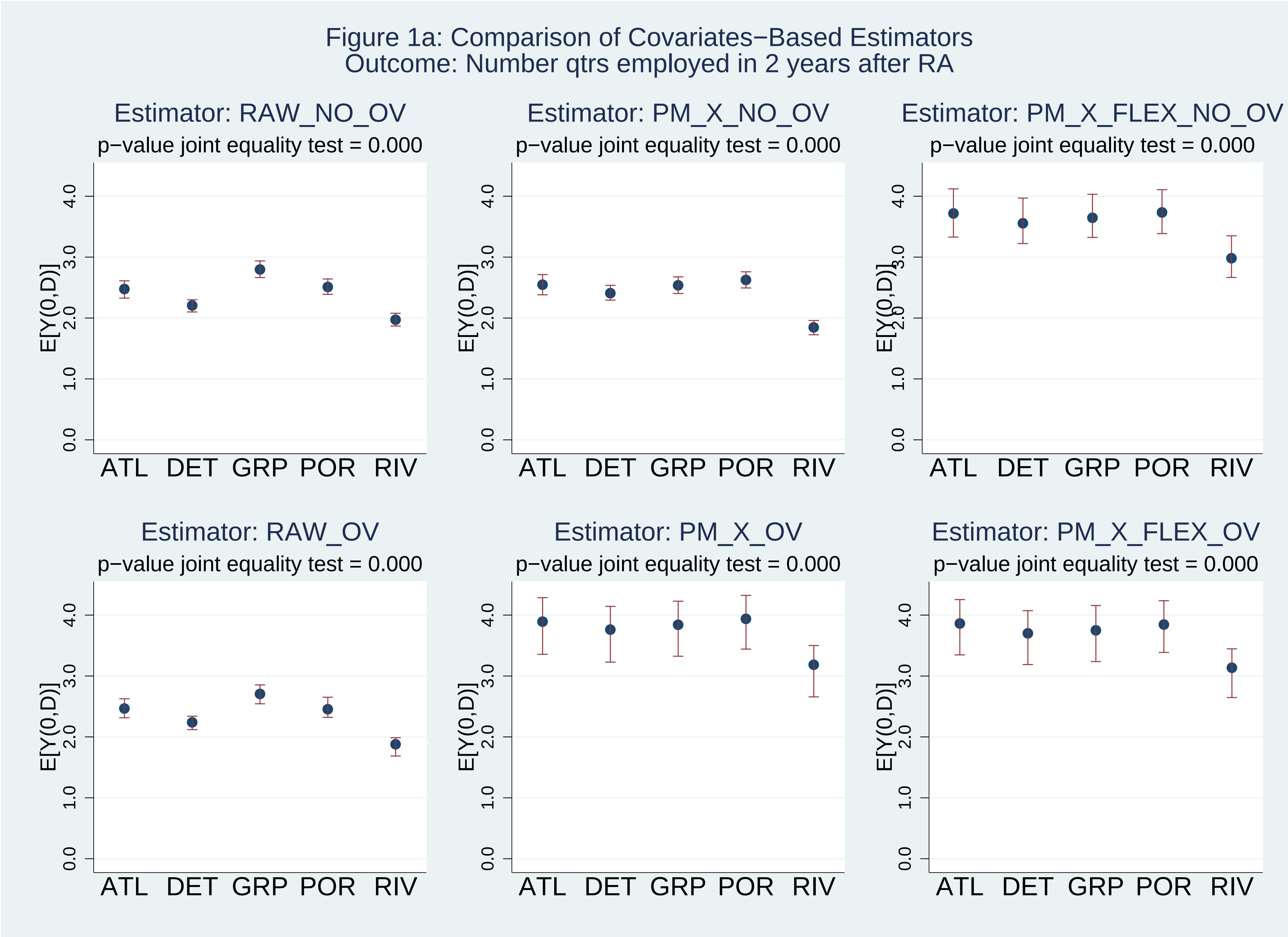
Figure 1a: Comparison of Covariates−Based Estimators
Outcome: Number qtrs employed in 2 years after RA
Estimator: RAW_NO_OV
p−value joint equality test = 0.000
Estimator: PM_X_NO_OV
p−value joint equality test = 0.000
Estimator: PM_X_FLEX_NO_OV
p−value joint equality test = 0.000
Estimator: RAW_OV
p−value joint equality test = 0.000
Estimator: PM_X_OV
p−value joint equality test = 0.000
Estimator: PM_X_FLEX_OV
p−value joint equality test = 0.000
E[Y(0,D)]
0.0
1.0
2.0
3.0
4.0
ATL
DET
GRP
POR
RIV

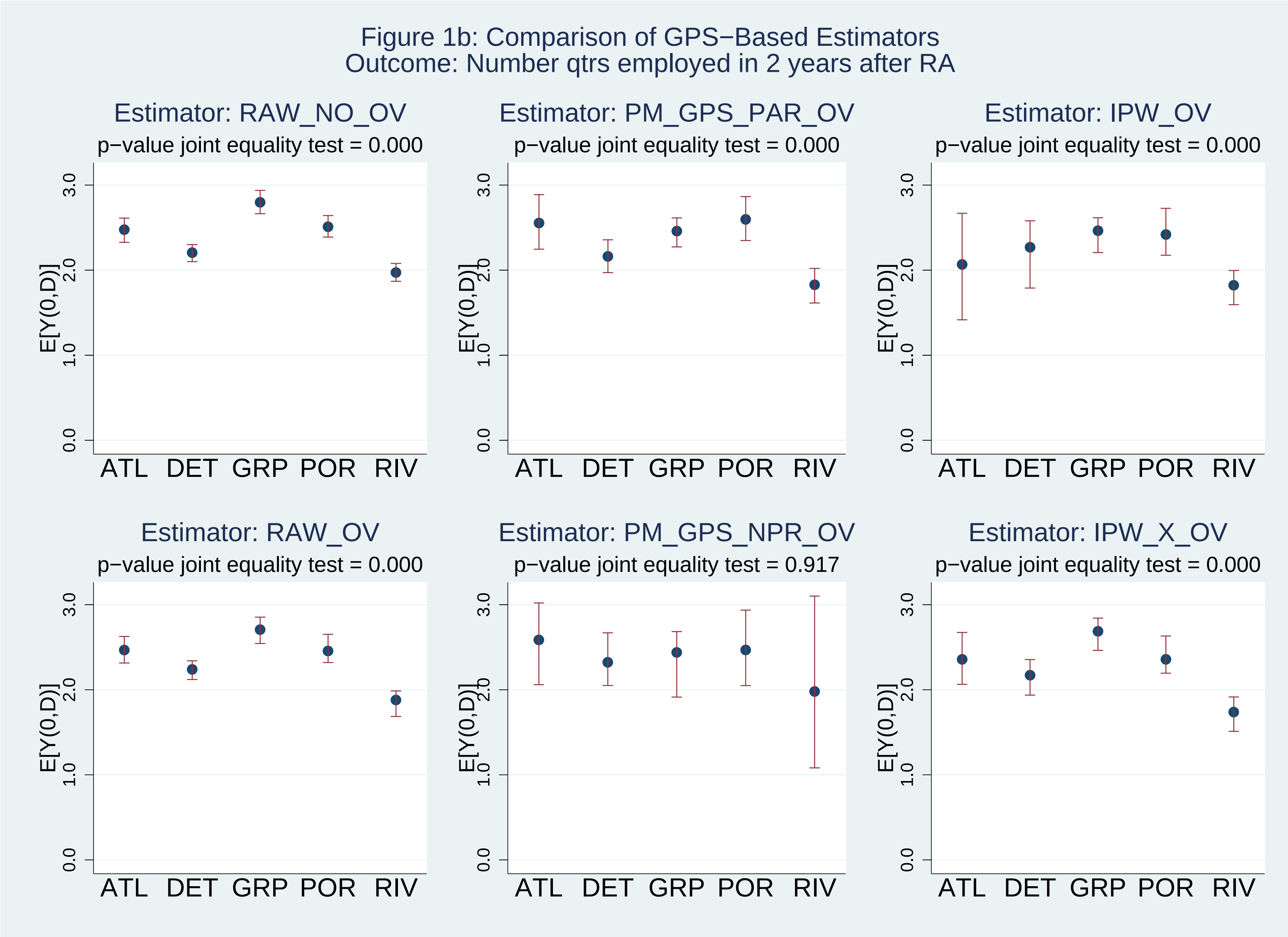

Figure 1b: Comparison of GPS−Based Estimators
Outcome: Number qtrs employed in 2 years after RA
Estimator: RAW_NO_OV
p−value joint equality test = 0.000
Estimator: PM_GPS_PAR_OV
p−value joint equality test = 0.000
Estimator: IPW_OV
p−value joint equality test = 0.000
Estimator: RAW_OV
p−value joint equality test = 0.000
Estimator: PM_GPS_NPR_OV
p−value joint equality test = 0.917
Estimator: IPW_X_OV
p−value joint equality test = 0.000
E[Y(0,D)]
0.0
1.0
2.0
3.0
ATL
DET
GRP
POR
RIV

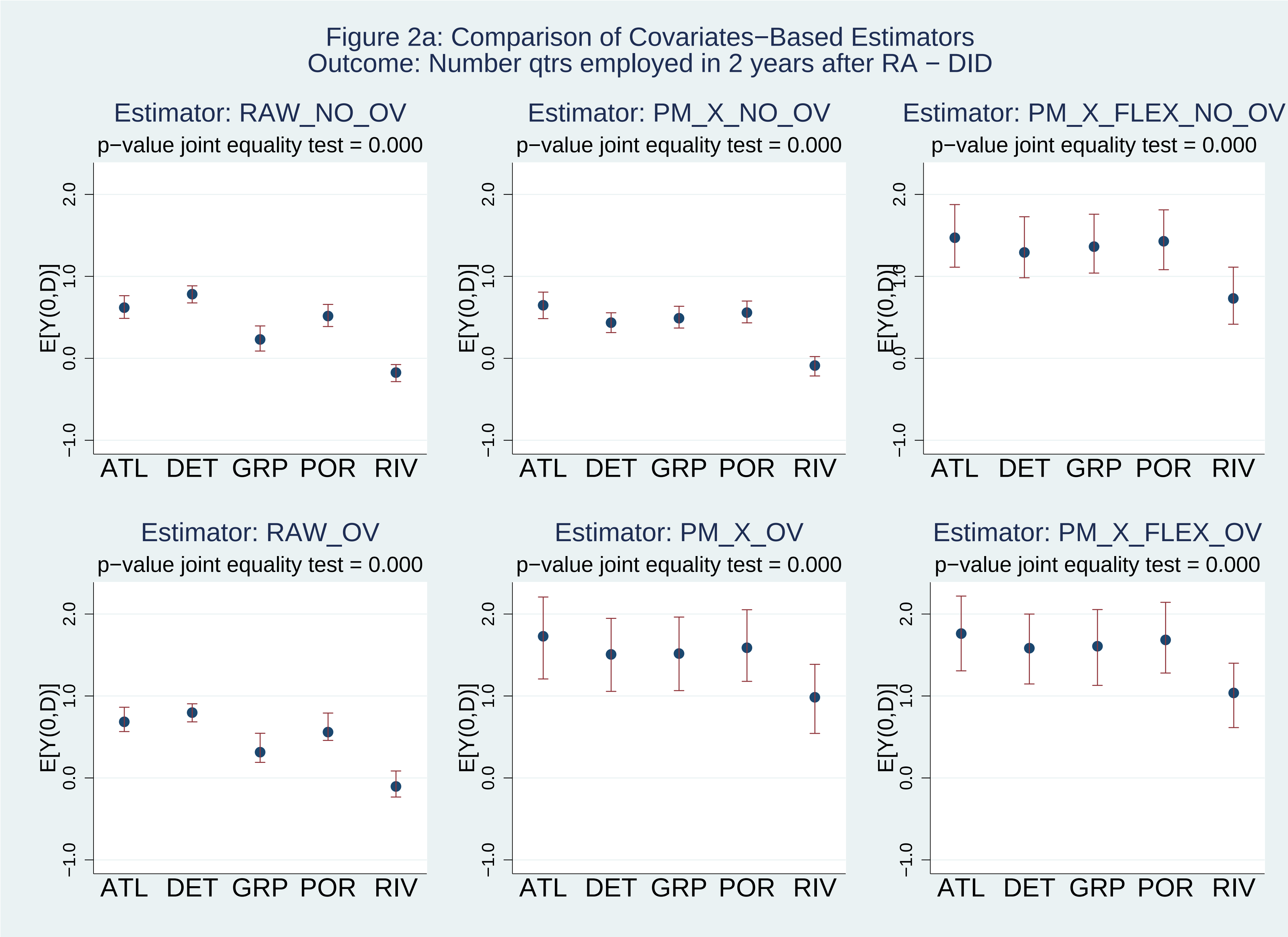
Figure 2a: Comparison of Covariates−Based Estimators
Outcome: Number qtrs employed in 2 years after RA − DID
Estimator: RAW_NO_OV
p−value joint equality test = 0.000
Estimator: PM_X_NO_OV
p−value joint equality test = 0.000
Estimator: PM_X_FLEX_NO_OV
p−value joint equality test = 0.000
Estimator: RAW_OV
p−value joint equality test = 0.000
Estimator: PM_X_OV
p−value joint equality test = 0.000
Estimator: PM_X_FLEX_OV
p−value joint equality test = 0.000
E[Y(0,D)]
−1.0
0.0
1.0
2.0
ATL
DET
GRP
POR
RIV

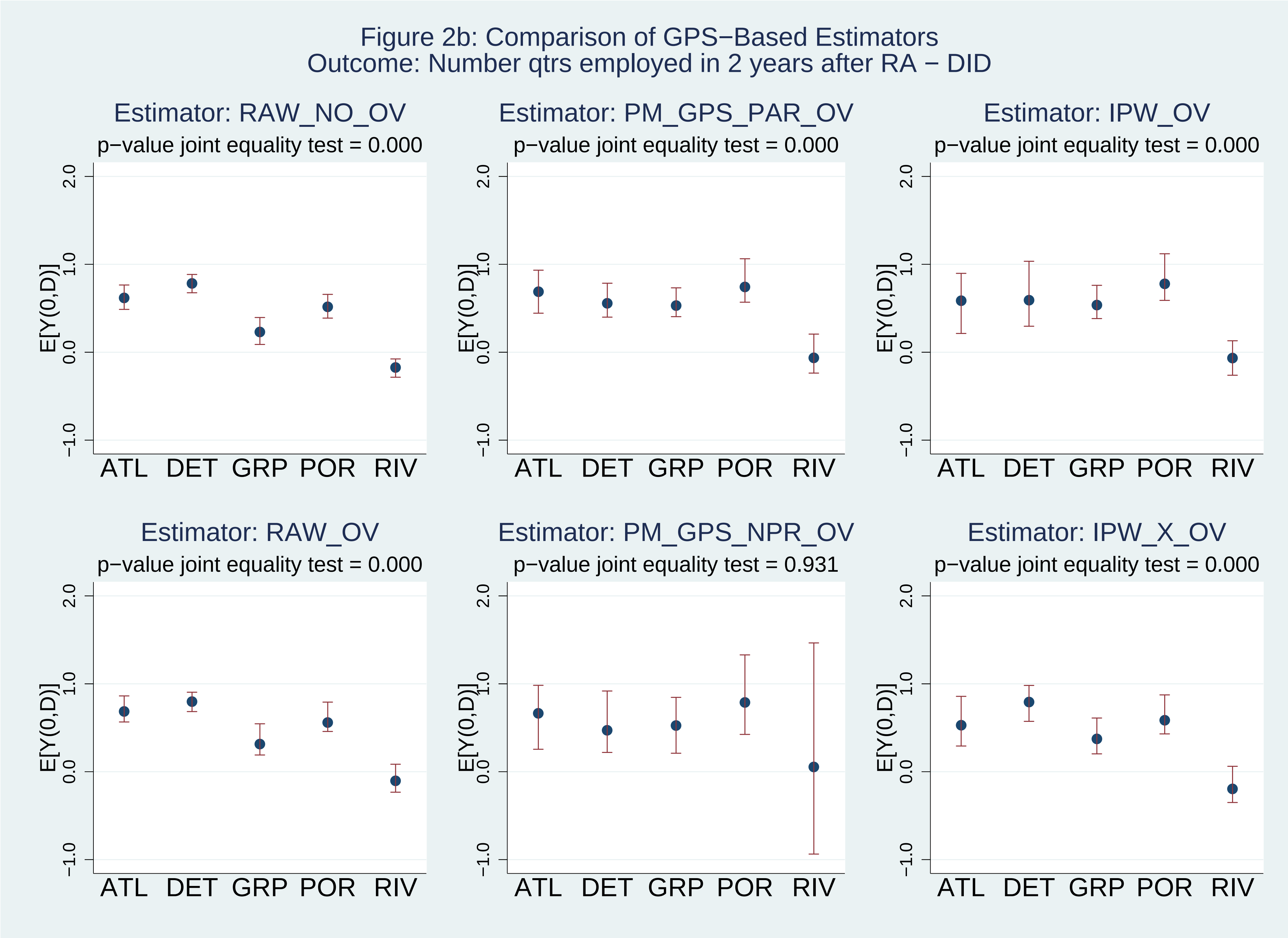
Figure 2b: Comparison of GPS−Based Estimators
Outcome: Number qtrs employed in 2 years after RA − DID
Estimator: RAW_NO_OV
p−value joint equality test = 0.000
Estimator: PM_GPS_PAR_OV
p−value joint equality test = 0.000
Estimator: IPW_OV
p−value joint equality test = 0.000
Estimator: RAW_OV
p−value joint equality test = 0.000
Estimator: PM_GPS_NPR_OV
p−value joint equality test = 0.931
Estimator: IPW_X_OV
p−value joint equality test = 0.000
E[Y(0,D)]
−1.0
0.0
1.0
2.0
ATL
DET
GRP
POR
RIV

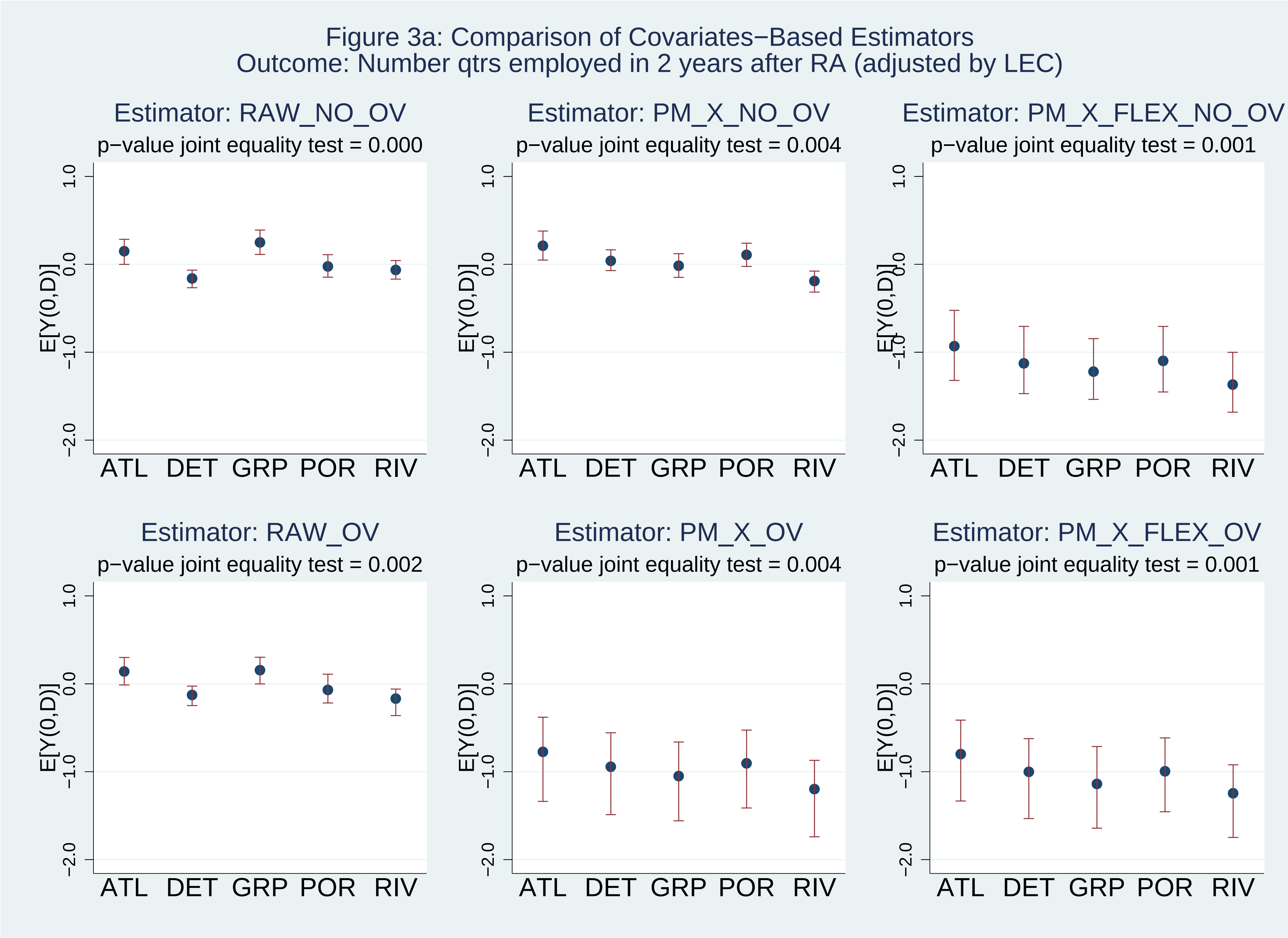

Figure 3a: Comparison of Covariates−Based Estimators
Outcome: Number qtrs employed in 2 years after RA (adjusted by LEC)
Estimator: RAW_NO_OV
p−value joint equality test = 0.000
Estimator: PM_X_NO_OV
p−value joint equality test = 0.004
Estimator: PM_X_FLEX_NO_OV
p−value joint equality test = 0.001
Estimator: RAW_OV
p−value joint equality test = 0.002
Estimator: PM_X_OV
p−value joint equality test = 0.004
Estimator: PM_X_FLEX_OV
p−value joint equality test = 0.001
E[Y(0,D)]
1.0
0.0
−1.0
−2.0
ATL
DET
GRP
POR
RIV

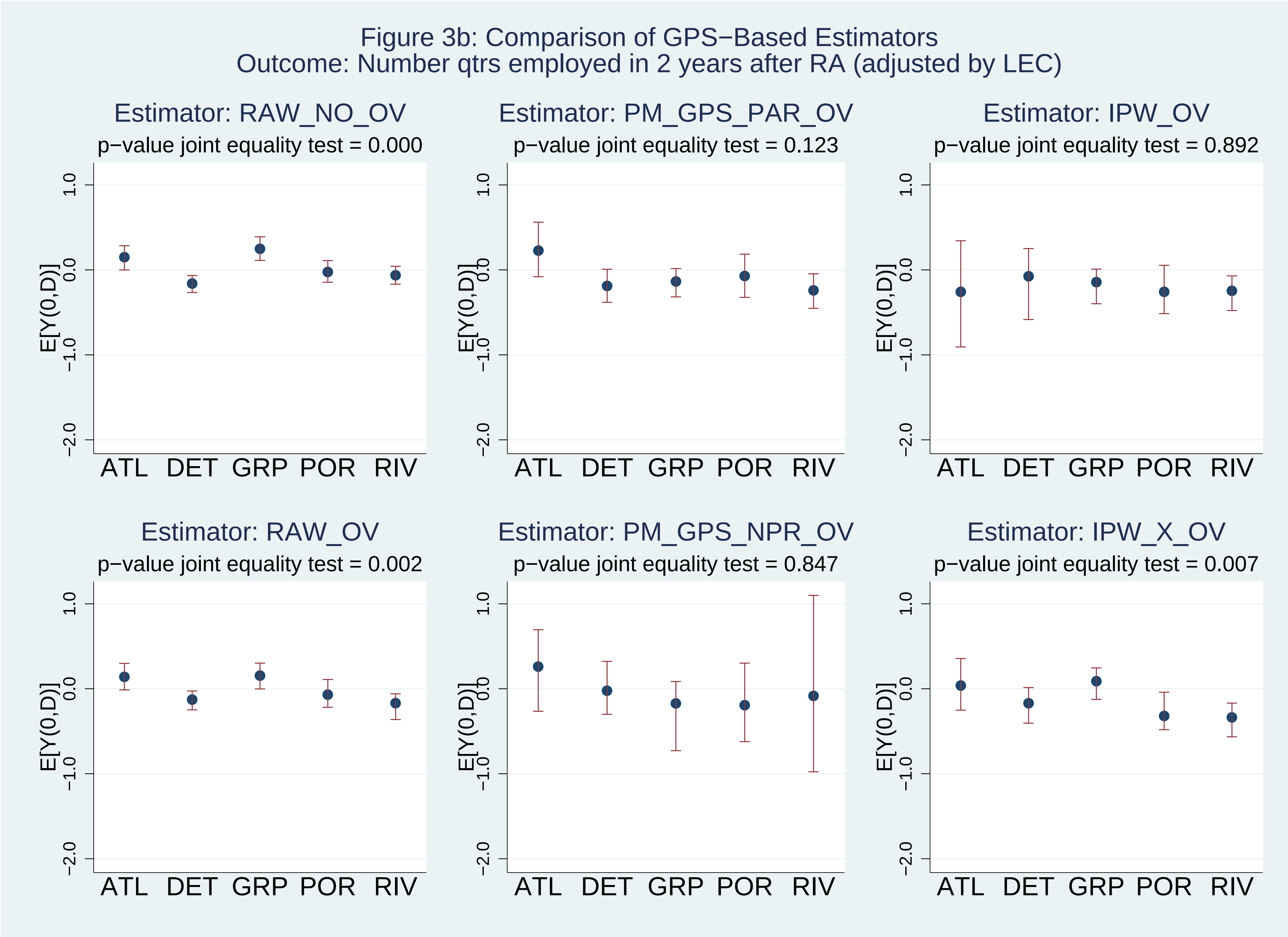
Figure 3b: Comparison of GPS−Based Estimators
Outcome: Number qtrs employed in 2 years after RA (adjusted by LEC)
Estimator: RAW_NO_OV
p−value joint equality test = 0.000
Estimator: PM_GPS_PAR_OV
p−value joint equality test = 0.123
Estimator: IPW_OV
p−value joint equality test = 0.892
Estimator: RAW_OV
p−value joint equality test = 0.002
Estimator: PM_GPS_NPR_OV
p−value joint equality test = 0.847
Estimator: IPW_X_OV
p−value joint equality test = 0.007
E[Y(0,D)]
1.0
0.0
−1.0
−2.0
ATL
DET
GRP
POR
RIV

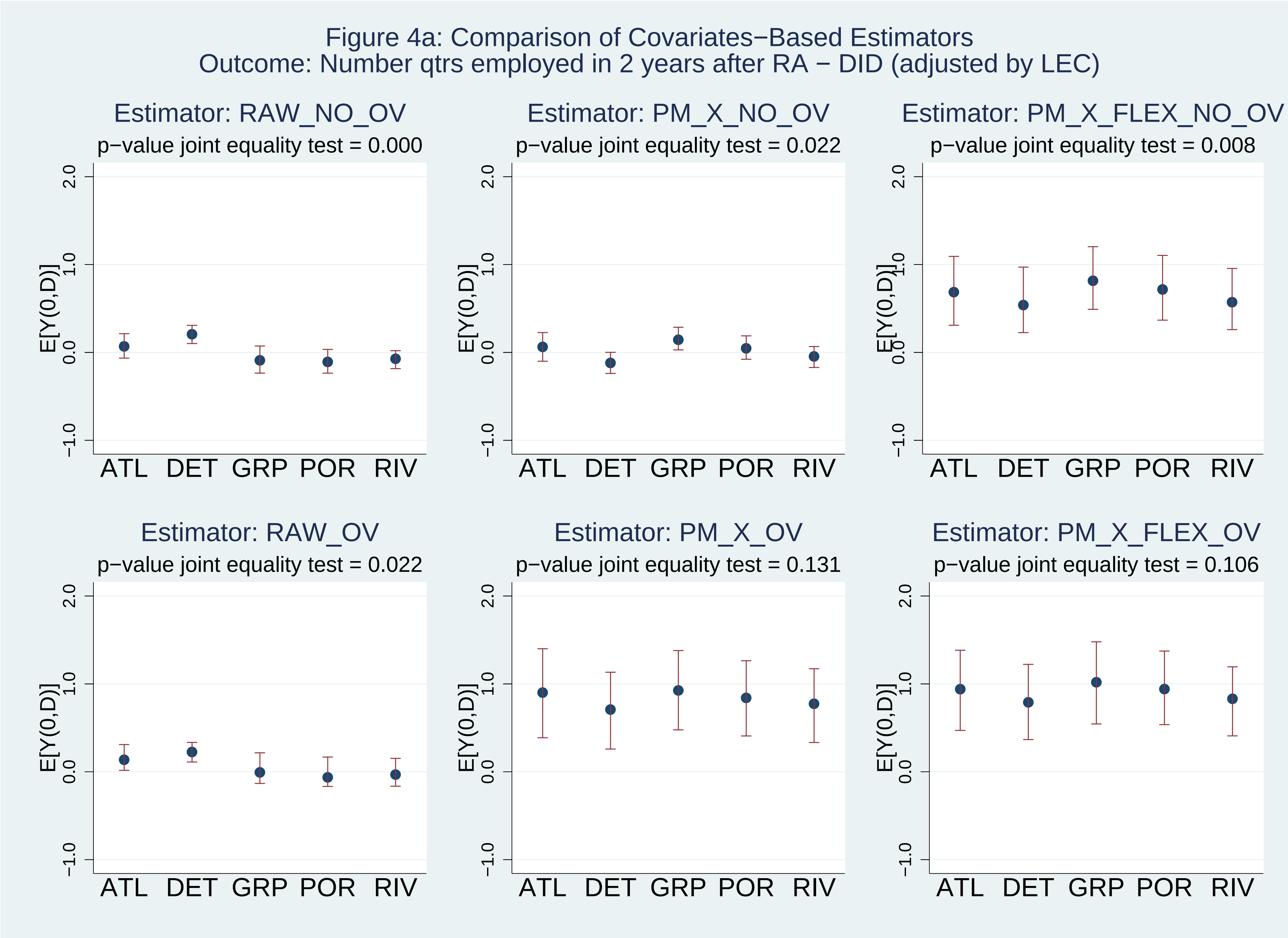
Figure 4a: Comparison of Covariates−Based Estimators
Outcome: Number qtrs employed in 2 years after RA − DID (adjusted by LEC)
Estimator: RAW_NO_OV
p−value joint equality test = 0.000
Estimator: PM_X_NO_OV
p−value joint equality test = 0.022
Estimator: PM_X_FLEX_NO_OV
p−value joint equality test = 0.008
Estimator: RAW_OV
p−value joint equality test = 0.022
Estimator: PM_X_OV
p−value joint equality test = 0.131
Estimator: PM_X_FLEX_OV
p−value joint equality test = 0.106
E[Y(0,D)]
−1.0
0.0
1.0
2.0
ATL
DET
GRP
POR
RIV

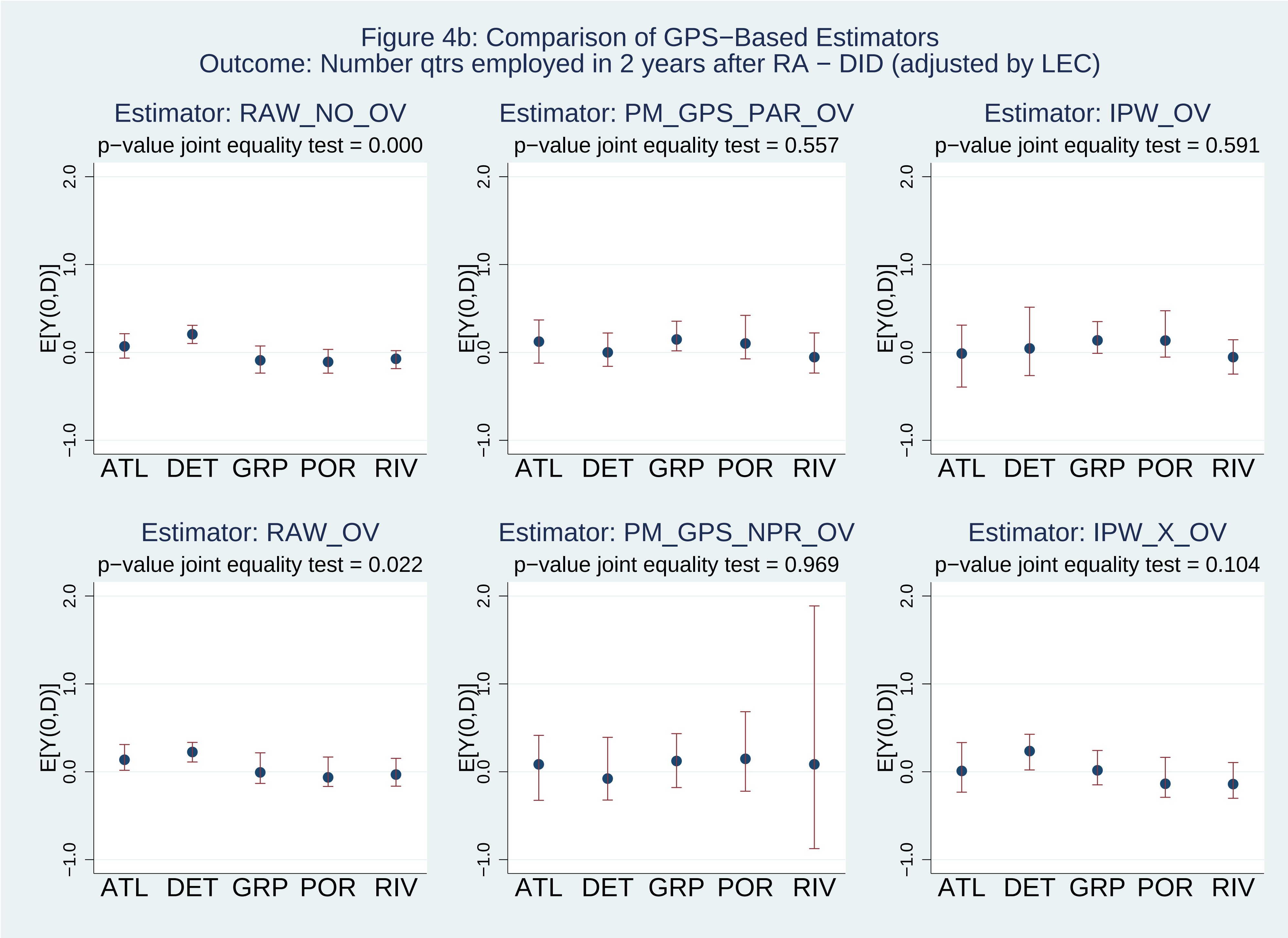
Figure 4b: Comparison of GPS−Based Estimators
Outcome: Number qtrs employed in 2 years after RA − DID (adjusted by LEC)
Estimator: RAW_NO_OV
p−value joint equality test = 0.000
Estimator: PM_GPS_PAR_OV
p−value joint equality test = 0.557
Estimator: IPW_OV
p−value joint equality test = 0.591
Estimator: RAW_OV
p−value joint equality test = 0.022
Estimator: PM_GPS_NPR_OV
p−value joint equality test = 0.969
Estimator: IPW_X_OV
p−value joint equality test = 0.104
E[Y(0,D)]
−1.0
0.0
1.0
2.0
ATL
DET
GRP
POR
RIV

Appendix Table 1. Balancing of Covariates Analysis

Variable	P-Values Joint tests of equality of means of covariates across all sites			Standardized differences of means of covariates in one site vs all other sites pooled together														
				Means Before Overlap					Means After Overlap					Means after Blocking on GPS				
	Raw	Raw w/Ovlp	GPS IPW	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV	ATL	DET	GRP	POR	RIV
Black	0.000	0.000	0.040	1.10***	1.04***	-0.17	-0.70	-0.89	0.96***	0.90***	-0.21	-0.80	-0.81	0.43***	0.21***	-0.09	-0.15	-0.16
Age 30-39 years old	0.000	0.000	0.101	0.24***	-0.14	-0.26	-0.02	0.14***	0.25***	-0.14	-0.21	0.00	0.10***	0.00	-0.05	-0.04	0.06	0.00
Age 40+ years old	0.000	0.000	0.903	0.09***	0.01	-0.10	-0.12	0.09***	0.07**	-0.01	-0.07	-0.11	0.09***	0.01	-0.01	-0.01	-0.03	-0.03
Teenage mother	0.000	0.000	0.513	0.11***	0.10***	0.24***	-0.17	-0.17	0.08***	0.10***	0.20***	-0.18	-0.15	0.03	0.02	0.04	-0.11	0.00
Never married	0.000	0.000	0.169	0.24***	0.44***	0.14***	-0.07	-0.53	0.17***	0.37***	0.08***	-0.10	-0.47	0.10*	0.10**	-0.01	-0.07	-0.10
Any child 0-5 years old	0.000	0.000	0.028	-0.47	0.11***	0.19***	0.24***	-0.09	-0.43	0.12***	0.14***	0.20***	-0.06	-0.01	0.03	0.00	0.04	0.05
Any child 6-12 years old	0.000	0.000	0.129	0.37***	-0.18	-0.28	-0.06	0.13***	0.36***	-0.16	-0.23	-0.04	0.09***	0.01	-0.09	-0.04	-0.02	-0.02
2 children in household	0.004	0.011	0.709	0.03	-0.08	0.08***	0.02	-0.01	0.02	-0.08	0.08**	0.02	-0.01	0.05	-0.01	0.01	0.00	0.00
3+ children in household	0.000	0.000	0.235	0.10***	-0.01	-0.23	0.07**	0.04*	0.11***	0.00	-0.19	0.05	0.01	-0.05	0.01	-0.05	0.04	0.01
10th grade	0.000	0.001	0.851	0.01	0.03	-0.01	0.13***	-0.12	-0.02	0.01	-0.02	0.11***	-0.08	-0.05	0.03	-0.01	-0.01	-0.03
11th grade	0.000	0.000	0.868	-0.09	0.16***	-0.01	0.04	-0.10	-0.08	0.16***	0.00	0.00	-0.11	-0.02	0.01	-0.01	0.01	-0.03
Grade 12 or higher	0.000	0.000	0.539	0.10***	-0.08	0.03	-0.20	0.13***	0.14***	-0.05	0.02	-0.17	0.07***	0.12**	-0.01	0.01	0.01	0.07*
Highest degree = HS/GED	0.000	0.002	0.477	-0.03	-0.15	0.01	-0.01	0.14***	0.02	-0.10	0.01	0.00	0.07***	0.02	-0.06	0.00	-0.02	0.04
Lives public/subss house	0.000	0.000	0.163	1.12***	-0.43	-0.14	0.26***	-0.42	1.06***	-0.47	-0.13	0.19***	-0.37	0.00	0.09*	-0.03	0.00	-0.09
1-2 moves in past 2 years	0.000	0.008	0.891	-0.03	-0.05	0.02	-0.07	0.10***	-0.03	-0.03	0.05*	-0.07	0.07***	0.02	0.06	0.01	0.04	0.05
3+ moves in past 2 years	0.000	0.000	0.072	-0.29	-0.33	0.24***	0.17***	0.18***	-0.24	-0.27	0.19***	0.14***	0.20***	-0.11	-0.07	0.02	0.09**	0.06*
On welfare < 2 years	0.000	0.000	0.011	-0.19	-0.29	0.12***	-0.03	0.30***	-0.16	-0.21	0.09***	0.00	0.26***	-0.06	-0.01	0.03	0.04	0.07**
On welfare for 2-5 years	0.000	0.000	0.206	-0.09	-0.10	0.05*	0.17***	-0.02	-0.11	-0.11	0.08**	0.17***	0.00	-0.06	0.01	0.01	0.02	0.03
On welfare 5-10 years	0.000	0.000	0.979	0.11***	0.11***	-0.09	0.07**	-0.15	0.10***	0.09***	-0.09	0.05*	-0.15	0.07	-0.01	-0.03	-0.04	-0.03
On Welfare Q1 before RA	0.000	0.000	0.454	0.45***	0.28***	-0.16	-0.10	-0.33	0.31***	0.14***	-0.10	-0.06	-0.26	0.08	0.08*	-0.07	-0.10	-0.02
On Welfare Q2 before RA	0.000	0.000	0.623	0.56***	0.41***	-0.04	0.08***	-0.69	0.38***	0.21***	-0.07	0.05**	-0.49	0.11**	0.13***	-0.10	-0.11	-0.04
On Welfare Q3 before RA	0.000	0.000	0.798	0.42***	0.44***	0.01	0.09***	-0.67	0.27***	0.27***	-0.04	0.06**	-0.49	0.16***	0.12***	-0.07	-0.11	-0.06
On Welfare Q4 before RA	0.000	0.000	0.821	0.21***	0.48***	0.06**	0.11***	-0.63	0.15***	0.33***	0.00	0.07***	-0.51	0.13**	0.10**	-0.05	-0.08	-0.07
On Welfare Q5 before RA	0.000	0.000	0.807	0.18***	0.51***	0.07**	0.07***	-0.61	0.12***	0.37***	0.01	0.03	-0.50	0.15***	0.10**	-0.05	-0.07	-0.06
On Welfare Q6 before RA	0.000	0.000	0.498	0.17***	0.52***	0.05*	0.05*	-0.59	0.12***	0.40***	0.00	0.01	-0.50	0.18***	0.09**	-0.03	-0.06	-0.03
On Welfare Q7 before RA	0.000	0.000	0.599	0.18***	0.53***	-0.01	0.04	-0.56	0.13***	0.42***	-0.04	0.00	-0.49	0.18***	0.10**	-0.03	-0.04	-0.03
Rec. FS in Q1 before RA	0.000	0.000	0.299	0.46***	0.40***	0.08***	0.13***	-0.74	0.25***	0.19***	0.03	0.08***	-0.47	0.14***	0.15***	-0.09	-0.13	-0.03
Rec. FS in Q2 before RA	0.000	0.000	0.345	0.59***	0.47***	0.09***	0.25***	-0.97	0.36***	0.23***	0.04	0.18***	-0.67	0.18***	0.19***	-0.11	-0.13	-0.05
Rec. FS in Q3 before RA	0.000	0.000	0.263	0.53***	0.49***	0.07**	0.25***	-0.93	0.33***	0.27***	0.01	0.19***	-0.68	0.20***	0.15***	-0.09	-0.13	-0.07
Rec. FS in Q4 before RA	0.000	0.000	0.418	0.41***	0.53***	0.13***	0.26***	-0.93	0.24***	0.33***	0.05*	0.20***	-0.71	0.15***	0.16***	-0.07	-0.10	-0.08
Rec. FS in Q5 before RA	0.000	0.000	0.480	0.36***	0.53***	0.10***	0.24***	-0.88	0.22***	0.36***	0.03	0.19***	-0.69	0.15***	0.11***	-0.06	-0.07	-0.08
Rec. FS in Q6 before RA	0.000	0.000	0.138	0.34***	0.53***	0.09***	0.24***	-0.86	0.22***	0.36***	0.02	0.18***	-0.69	0.18***	0.11**	-0.05	-0.06	-0.06
Rec. FS in Q7 before RA	0.000	0.000	0.071	0.34***	0.54***	0.05*	0.22***	-0.82	0.22***	0.39***	-0.01	0.16***	-0.68	0.17***	0.12***	-0.04	-0.05	-0.09
RA in year 1993	0.000	0.000	0.171	-0.21	0.19***	-0.15	0.77***	-0.49	-0.30	0.11***	-0.21	0.75***	-0.36	0.02	0.03	-0.06	0.11***	-0.04
Employed Q1 before RA	0.000	0.000	0.037	-0.10	-0.11	0.20***	0.04	0.00	-0.07	-0.07	0.20***	0.02	-0.04	-0.03	-0.09	0.05	0.01	-0.04
Employed Q2 before RA	0.000	0.000	0.209	-0.15	-0.15	0.19***	0.06**	0.06**	-0.11	-0.10	0.17***	0.06**	0.01	-0.02	-0.09	0.03	0.01	-0.02
Employed Q3 before RA	0.000	0.000	0.538	-0.12	-0.18	0.16***	0.04	0.09***	-0.10	-0.12	0.15***	0.05*	0.05**	0.01	-0.07	0.04	0.01	-0.01
Employed Q4 before RA	0.000	0.000	0.260	-0.06	-0.20	0.15***	-0.02	0.12***	-0.04	-0.14	0.14***	0.00	0.08***	0.01	-0.11	0.06*	0.01	0.01
Employed Q5 before RA	0.000	0.000	0.202	-0.01	-0.22	0.16***	-0.02	0.10***	-0.01	-0.18	0.14***	-0.01	0.09***	0.00	-0.08	0.05	0.03	0.02
Employed Q6 before RA	0.000	0.000	0.454	0.01	-0.24	0.21***	-0.03	0.08***	0.02	-0.19	0.19***	-0.02	0.07***	-0.01	-0.11	0.05	0.02	0.02
Employed Q7 before RA	0.000	0.000	0.084	0.04	-0.25	0.24***	-0.02	0.05**	0.04	-0.23	0.23***	0.00	0.05**	0.01	-0.08	0.06*	-0.04	0.04
Employed Q8 before RA	0.000	0.000	0.120	0.04	-0.29	0.28***	-0.01	0.06**	0.04	-0.26	0.24***	0.01	0.06**	0.02	-0.09	0.04	-0.04	0.04
Emply at RA (self reported)	0.000	0.000	0.013	-0.10	-0.13	0.12***	-0.06	0.13***	-0.07	-0.09	0.15***	-0.04	0.08***	-0.07	-0.02	0.06*	0.01	0.04
Ever wrkd FT 6+ mths sm. job	0.000	0.000	0.726	0.15***	-0.52	-0.04	0.28***	0.16***	0.17***	-0.47	-0.03	0.29***	0.14***	-0.11	-0.05	-0.03	0.09**	0.07*
Earnings Q1 before RA	0.000	0.000	0.148	-0.11	-0.15	0.05	0.01	0.15***	-0.07	-0.08	0.05*	0.02	0.08***	-0.02	-0.09	0.05	0.01	0.00
Earnings Q2 before RA	0.000	0.000	0.195	-0.17	-0.20	0.08***	-0.03	0.23***	-0.10	-0.11	0.09***	-0.01	0.12***	-0.01	-0.09	0.05	0.01	0.02
Earnings Q3 before RA	0.000	0.000	0.381	-0.17	-0.21	0.07**	-0.06	0.27***	-0.11	-0.12	0.07**	-0.03	0.17***	-0.01	-0.08	0.05	0.01	0.02
Earnings Q4 before RA	0.000	0.000	0.617	-0.07	-0.24	0.03	-0.06	0.26***	-0.03	-0.16	0.06*	-0.03	0.17***	0.00	-0.09	0.05	0.02	0.01
Earnings Q5 before RA	0.000	0.000	0.104	-0.03	-0.23	0.02	-0.07	0.24***	-0.01	-0.16	0.05*	-0.04	0.17***	0.01	-0.11	0.06*	0.00	0.03
Earnings Q6 before RA	0.000	0.000	0.208	0.03	-0.24	0.03	-0.06	0.21***	0.03	-0.18	0.05	-0.04	0.16***	-0.01	-0.09	0.05	0.02	0.05
Earnings Q7 before RA	0.000	0.000	0.106	0.07**	-0.25	0.04	-0.07	0.19***	0.06**	-0.20	0.06**	-0.04	0.15***	-0.02	-0.07	0.05	-0.01	0.06
Earnings Q8 before RA	0.000	0.000	0.063	0.07**	-0.26	0.04	-0.06	0.19***	0.06**	-0.21	0.06*	-0.03	0.16***	0.00	-0.09	0.04	-0.03	0.05
Any earns yr before RA (slf-rep)	0.000	0.000	0.015	-0.26	-0.36	0.31***	0.08***	0.20***	-0.18	-0.28	0.30***	0.07**	0.15***	-0.07	-0.17	0.07**	0.08*	0.05