Evaluating the Effect of Training on Wages in the Presence of Noncompliance, Nonemployment, and Missing Outcome Data

Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data

Paolo Frumento, Karolinska Institutet, paolo.frumento@ki.se

Fabrizia Mealli, Università di Firenze, mealli@ds.unifi.it

Barbara Pacini, Università di Pisa, barbara.pacini@sp.unipi.it

Donald B. Rubin, Harvard University, rubin@stat.harvard.edu

Abstract: The effects of a job-training program on both employment and wages are evaluated, using data from a randomized study. Principal stratification is used to address, simultaneously, the complications of noncompliance, wages that are only partially defined because of nonemployment, and unintended missing outcomes. The first two complications are of substantive interest, whereas the third is a nuisance. The objective is to find a parsimonious model that can be used to inform public policy. We conduct a likelihood-based analysis using finite mixture models estimated by the EM algorithm. We maintain an exclusion restriction assumption for the effect of assignment on employment and wages for noncompliers, but not on missingness. We provide estimates under the Missing at Random assumption, and assess the robustness of our results to deviations from it. The plausibility of meaningful restrictions is investigated by means of scaled log-likelihood ratio statistics. Substantive conclusions include the following. For compliers, the effect on employment is negative in the short term; it becomes positive in the long term, but these effects are small at best. For always employed compliers, i.e., compliers who are employed whether trained or not trained, positive effects on wages are found at all time periods. Our analysis reveals that background characteristics of individuals differ markedly across the principal strata. We found evidence that the program should have been better targeted, in the sense of being designed differently for different groups of people, and specific suggestions are offered. Previous analyses of this data set, which did not address all complications in a principled manner, led to less nuanced conclusions about Job Corps.

Keywords: EM algorithm, finite mixture models, missing at random, noncompliance, partially defined outcomes, principal stratification, Rubin causal model, training, wages.

1 The Job Corps Study and its complications

Evaluations of government-sponsored job-training programs have typically been undertaken using data from nonrandomized studies (e.g., Dehejia and Wahba, 1999). Some social experiments have also been conducted, because a perfect randomized experiment is the generally accepted tool to infer causal effects, although this topic has been the subject of debate in the economic literature (e.g., Heckman *et al.*, 1999; Deaton, 2010; Heckman, 2010; Imbens, 2010).

In a randomized experiment, units are randomly assigned to the treatment group or to the control group, which ensures that treated and control units have the same expected distribution of all pre-randomization individual characteristics. Experiments, however, and social experiments in particular, often suffer from a number of substantially relevant complications, most notably noncompliance with assigned treatment and *partially defined outcomes* (e.g., quality of life when dead, called *truncation by death*, Rubin, 2000, 2006; McConnell *et al.*, 2008), as well as unintended missing outcomes. The presence of such complications can shift the focus to causal estimands that differ from the ones the experiment was originally designed to address, but the randomization still allows one to estimate some original causal effects for specific subgroups.

Here, we evaluate the effects of Job Corps, which is the largest, most comprehensive US education and job training program for disadvantaged youths between the ages of 16 and 24, using data from a randomized study, the National Job Corps Study, conducted by Mathematica Policy Research, Inc., involving a national random sample of all eligible applicants in late 1994 and 1995. Sampled youths were assigned randomly to the program (treatment) group ($9,409$) or the control group (5977), which was essentially embargoed from the program for three years. Interviews were planned at three subsequent points in time: 52, 130, and 208 weeks after random assignment. We focus on the effect of the program on employment and wages at these specific weeks.

Regarding the three complications, first, compliance with assigned treatment was imperfect,

with only 68% of those assigned to the program group immediately enrolling (within the first semester after assignment) and participating in the program for at least 1 week. Second, wages are not well defined for those who are not employed. Third, outcome variables are missing for some participants in the study at the various weeks. Most previous analyses of these data ignored noncompliance by focusing on intention-to-treat effects of being offered participation in Job Corps (Lee, 2009; Flores-Lagunes et al., 2007; Flores and Flores-Lagunes, 2009; Zhang et al., 2009); an exception is Schochet (2001), who estimated average effects for compliers using the standard econometric IV estimator of LATE (Imbens and Angrist, 1994), also called the Complier Average Causal Effect (CACE, Imbens and Rubin, 1997). However, all previous analyses dropped units with unintended missing outcomes, restricting the analysis to the subsample of individuals who both completed the 208-week interview and had no missing relevant outcome values on employment and wages (Burghardt et al., 2003; Lee, 2009; Zhang et al., 2009; Flores and Flores-Lagunes, 2009), thereby implicitly assuming that the missing data mechanism was Missing Completely at Random (MCAR; Little and Rubin, 2002).

In our analysis, we include all randomized units and explicitly address all three complications. The framework we adopt uses potential outcomes to define causal effects regardless of the mode of inference, often referred to as the Rubin Causal Model (RCM; Holland, 1986); causal effects are defined by comparisons of potential outcomes on a common set of units (Rubin, 1974, 1978, 2005). We apply Principal Stratification (PS; Frangakis and Rubin, 2002), which was originally introduced to address post-treatment complications within the RCM. PS has often been used to represent and solve single complications, but few papers have dealt with more than one complication simultaneously. In general, the analysis is more complicated than that in the presence of each of the complications separately, and complexity tends to grow exponentially with the number of distinct complications.

The three complications have different implications for causal analysis. Specifically, noncompliance and partially defined outcomes limit meaningful estimands to particular principal strata, and so require new definitions of causal estimands. Missing outcomes are tackled within the same general framework, but, in contrast, the parameters governing the missingness are considered to

be nuisance parameters. Because access to Job Corps was essentially denied to those assigned to the control group, we define compliers to be those individuals who would have immediately enrolled if offered the program and all others noncompliers. We further classify the individuals into principal strata according to the joint values, when assigned to be trained and when assigned not to be trained, of their (1) potential compliances, (2) potential employment statuses, and (3) potential missingness behaviors.

Our primary causal estimands are: the average causal effect on employment for compliers and the average causal effect on wages for the always-employed compliers (i.e., those compliers who would be employed irrespective of treatment assignment), and those estimands within subgroups defined by observed characteristics. Other policy-relevant estimands are the relative sizes of the various principal strata and, when employed, the within-principal-stratum distributions of wages, under treatment or under control, as well as the distribution of covariates within principal strata.

We proceed as follows. Section 2 presents some descriptive statistics, showing how some naive conclusions regarding the effects of the program can be misleading. Section 3 discusses the framework needed to address the three complications simultaneously. Section 4 outlines the likelihood approach used to characterize the effects of Job Corps. Section 5 presents the results of the empirical analysis and provides some concluding remarks.

2 General considerations and descriptive univariate summaries

For all units in the National Job Corps Study, covariates, variables unaffected by treatment assignment, were collected ($\mathbf{X}$). Some subpopulations defined by $\mathbf{X}$ were randomized into the program versus control group with varying, but known, probabilities. We use all units from the original research sample: we eliminated only the few units who did not complete the baseline interview, the units who died during the follow-up, and the units who, although assigned to the control group, were admitted to the program and excluded from the Job Corps study.

Summary statistics for many of the covariates used in our analysis are displayed in Table 1 ($N = 13,987$), and reveal some missing values. We addressed this missing data problem using

the MICE (van Buuren and Oudshoorn, 1999) procedure in R to multiply impute the incomplete multivariate data. We used only the baseline covariates as predictors in the chained equations, and included as fully observed covariates indicators of missingness for each of the covariates with a percentage of missing values above 20%. Linear regressions were used for numerical covariates; binary/multinomial logistic models for dichotomous/polytomous variables. Ten different imputations were generated; given the very small variability of results across multiple imputations, as found also in Zhang *et al.* (2009), we only present the results from one singly imputed data set (Rubin, 1987).

Table 2 presents summary statistics for compliance and outcome variables: employment, total earnings, weekly hours, and hourly wages at the three follow-up interviews. We use a single missing data indicator at each time point because hours worked (employment status) and wages are nearly always either both observed or both missing. There is actually a small number of units (38, 34, 60 at the three weeks, respectively) for whom the employment status is observed but information on wages is partially missing (e.g., wages are known only for some jobs but not for others). For these situations, wages were constructed using the same imputation strategy used by Schochet (2001) in the publicly released data for the subsample of respondents to the 208-week interview.

Some naive conclusions about the effects of the training program can be drawn from Table 2. For example, by comparing the employment rate of respondent treated units with that of respondent control units, we observe a negative effect at week 52 (−6%), and a small but positive effect at weeks 130 and 208 (2% and 4%, respectively). These contrasts can be formally interpreted, however, neither as estimates of the effect of participation in the program, because they neglect noncompliance, nor as estimates of ITT effects, because they also neglect nonresponse, unless under the implausible MCAR assumption. Similarly, we can naively compare the average hourly wage of respondent employed treated units with that of respondent employed control units, showing positive effects on wages ranging from 0.24 $/hour to 0.34 $/hour. Again, these estimates neglect missing data, noncompliance, and partially defined wages due to nonemployment, thus contrasting averages in groups that are not comparable. Some additional informal comparisons can

be computed in the form of moment-based IV estimates of local average treatment effects (LATEs, Imbens and Angrist, 1994), as the ratio of ITT effects for outcomes (i.e., comparisons by treatment assignment) to the proportion of compliers, still neglecting missing data. For employment, these estimated LATEs are equal to −0.10, 0.03, and 0.06 at week 52, 130 and 208, respectively. In the case of wages, estimated LATEs were computed using only the employed units, thus neglecting missing outcomes and partially defined wages; they are 0.43, 0.48, and 0.34 $/hour at the three weeks respectively.

3 Technical framework

3.1 General setup

Consider a large hypothetical super-population of individuals, each of whom can potentially participate in Job Corps and be assigned treatment z, with $z = 1$ for active treatment (i.e., offered enrollment in Job Corps), $z = 0$ for control. A probability sample of N individuals from this super-population comprises the participants in the study.

We adopt the Rubin Causal Model as a framework to define causal effects. Assuming SUTVA (Stable Unit Treatment Value Assumption; Rubin, 1978, 1980, 1990), we define, for each unit i and each post-treatment variable, two potential outcomes, each associated with one of the two treatment levels that unit i can potentially receive. SUTVA states that potential outcomes for individual i are unaffected by the treatment assignments of other individuals (no interference), and, for each unit, that there are no hidden versions of treatment or control being considered.

It is, in principle, feasible to use the PS framework to analyze the three weeks jointly; this, however, would imply a far larger number of possible principal strata and a consequent growing complexity of model specification and inference (as in Jin and Rubin, 2009). Although we consider each of the three weeks independently, our model is still quite complicated. Compared to analyses conducted by others, however, ours allows one to compare results over the three weeks under study, because all three are derived using all randomized units, without restriction to those with complete outcome data at the different weeks (available-case analysis). For unit i, we let $D_i(1)$

denote the binary compliance indicator when assigned treatment; i.e., $D_i(1) = 1$ implies unit i would immediately enroll in Job Corps if offered it, and $D_i(1) = 0$ otherwise; $D_i(0) = 0\ \forall i$ by definition, and so is suppressed notationally; $D_i(1) = 1, 0$ implies that the *ith* unit "does" or "does not" do as assigned. Compliance status does not change over time by definition and can be considered a covariate, which is not observed for units assigned to the control group. As for the other post-assignment variables, we suppress notation for the three outcome periods. Respectively, let $S_i(z)$, $W_i(z)$, and $M_i(z)$ represent the potential employment status indicators (1 = employed, 0 = nonemployed; "S" for "salaried"), the potential wages, and the potential missingness indicators, if individual i is assigned to treatment z, $z = 0, 1$. Following Zhang et al. (2008), because wages are well-defined only if $S_i(z) = 1$, we define the wages to be $W_i(z) = *$ when $S_i(z) = 0$. In our study, wages and employment status are either both observed or both missing, so that $M_i(z) = 0$ when $S_i(z)$ and $W_i(z)$ are both observed, and $M_i(z) = 1$ when $S_i(z)$ and $W_i(z)$ are both missing and are coded as "?". Individual causal effects are defined as comparisons of potential outcomes, e.g., $M_i(1) - M_i(0)$, $S_i(1) - S_i(0)$, and $W_i(1) - W_i(0)$, where this last quantity is defined to be $*$ if either $W_i(1)$ or $W_i(0)$ is $*$. At most three of the six potential outcomes are observed, those corresponding to the treatment level to which unit i is assigned.

The distribution of $\mathbf{Z}$ conditional on the observable potential outcomes and observed covariates defines the assignment mechanism, which allows us to draw inferences about causal estimands from the observed data. The random assignment of $\mathbf{Z}$ in our study within subpopulations defined by $\mathbf{X}$ means that:

$$Pr\left(\mathbf{Z}|\mathbf{D}(1), \mathbf{S}(0), \mathbf{S}(1), \mathbf{W}(0), \mathbf{W}(1), \mathbf{M}(0), \mathbf{M}(1), \mathbf{X}\right) = Pr\left(\mathbf{Z}|\mathbf{X}\right), \tag{1}$$

where the bold indicates column vectors of the corresponding unit indicators (e.g., $\mathbf{Z} = (Z_1, ..., Z_N)^T$). The observable potential outcomes for a sampled unit are a joint draw from the super-population distribution. Their distribution is, by definition, unit exchangeable, that is, invariant under a permutation of the unit indices. Therefore, appealing to deFinetti's theorem, with essentially no loss of generality, their joint distribution can be written as (Rubin, 1978):

$$f(\mathbf{D}(1), \mathbf{S}(0), \mathbf{S}(1), \mathbf{W}(0), \mathbf{W}(1), \mathbf{M}(0), \mathbf{M}(1), \mathbf{X}) = \tag{2}$$

$$\int \prod_i f(D_i(1), S_i(0), S_i(1), W_i(0), W_i(1), M_i(0), M_i(1)|\mathbf{X}_i, \boldsymbol{\theta}) f(\mathbf{X}_i|\boldsymbol{\varphi}) p(\boldsymbol{\theta}) p(\boldsymbol{\varphi}) d\boldsymbol{\theta} d\boldsymbol{\varphi},$$

where the global parameter $\boldsymbol{\theta}$ has prior distribution $p(\boldsymbol{\theta})$, and the parameter $\boldsymbol{\varphi}$ governing the distribution of $\mathbf{X}$, is a priori independent of $\boldsymbol{\theta}$. In what follows we will conduct a likelihood analysis for $\boldsymbol{\theta}$, assuming the value of $\boldsymbol{\theta}$ that governed the distribution of observable data has been drawn from a prior distribution with compact support.

In this framework, D, S, and M play the role of intermediate variables, which allow us to classify units into some principal strata, which are generally latent. Ignoring, for now, the missingness mechanism, units can be cross-classified by compliance status and employment status: $[\{c, n\} \times \{EE, EN, NE, NN\}]$, into eight groups, where we define:

- $c = \{i : D_i(1) = 1\}$, the subpopulation of compliers;
- $n = \{i : D_i(1) = 0\}$, the subpopulation of noncompliers;
- $EE = \{i : S_i(1) = S_i(0) = 1\}$, those who would be employed regardless of their treatment assignment; for this stratum, $W_i(1)$ and $W_i(0)$ are defined in $\Re^+$;
- $EN = \{i : S_i(1) = 1 \text{ and } S_i(0) = 0\}$, those who would be employed only if assigned treatment; for this stratum, $W_i(1) \in \Re^+$ and $W_i(0) = *$;
- $NE = \{i : S_i(1) = 0 \text{ and } S_i(0) = 1\}$, those who would be employed only if assigned to the control group; for this stratum, $W_i(1) = *$ and $W_i(0) \in \Re^+$;
- $NN = \{i : S_i(1) = S_i(0) = 0\}$, those who would be nonemployed regardless of their treatment assignment; for this stratum, $W_i(1) = W_i(0) = *$.

Without additional assumptions, group membership for unit i, $G_i = (D_i(1), S_i(0), S_i(1))$, which takes on values in $\{c\&EE, c\&EN, c\&NE, c\&NN, n\&EE, n\&EN, n\&NE, n\&NN\}$, is unobserved for all units; by the randomization, however, the eight types have, in expectation, the same distribution in both treatment groups. The strata can be considered covariates unaffected by treatment assignment, so that, in the same way randomization allows us to compare treated and control units with the same values of any X variable (e.g., females), we can compare treated and control units

belonging to the same principal stratum. The only complication is that principal strata can, in general, be only partially observed, so that conditioning on principal strata is not as simple as conditioning on fully observed covariates. Consequently, assumptions can play far more important roles with principal strata than with fully observed covariates. Some of these assumptions reduce the number of principal strata; others impose certain restrictions on the distribution of outcomes within or among strata: these include various forms of exclusion restrictions.

Without any assumptions on the missingness mechanism, each of the above eight principal strata is a mixture of four subgroups, according to the pair of potential missing indicators $M_i(1)$, $M_i(0)$ (units with outcomes never missing ($M_i(1) = 0$ and $M_i(0) = 0$), units with outcomes always missing ($M_i(1) = 1$ and $M_i(0) = 1$), units with outcomes missing only under control ($M_i(1) = 0$ and $M_i(0) = 1$), and units with outcomes missing only under treatment ($M_i(1) = 1$ and $M_i(0) = 0$)). Among the assumptions about the missing data process proposed in the literature, one that appears to be plausible in our context is Missing at Random (MAR; Rubin, 1976). MAR cannot be tested without auxiliary information, and its plausibility depends on the information available in a specific study. In general, MAR is more reasonable when the set of covariates contains rich information on units. MAR allows the missingness probabilities to depend on observed values but, given those, not on any missing values. If MAR holds and the parameters of the missing data mechanism are distinct from those of the outcome distribution, then the missing data process is said to be ignorable (Rubin, 1976), meaning that valid likelihood inference ignores the missing data model. The compliance indicator $D_i(1)$ is missing for 1% of units in the treatment group, presumably due to data coding errors. Throughout, we assume those indicators to be MAR, thus avoiding an explicit missingness indicator for them.

3.2 Assumptions and estimands

We assume exclusion restrictions for noncompliers for both W and S; that is, for noncompliers, potential outcomes do not depend on treatment assignment:

Exclusion restriction for S for noncompliers

If $D_i(z) = 0$ ($z = 0, 1$), then $S_i(0) = S_i(1)$.

Exclusion restriction for W for noncompliers

If $D_i(z) = 0$ $(z = 0, 1)$, then $W_i(0) = W_i(1)$.

These assumptions are substantive ones that may be violated depending on the empirical setting: Here, they appear rather plausible. The offer to be trained should not alter the activities or the labour market behavior of those units who are not willing to accept the offer within a reasonable length of time; in addition, potential employers are plausibly unaware of the assignment status of noncompliers, so that future job and wage offers cannot be affected by the assignment (see Angrist, 1990, and Angrist et al., 1996, for discussions of possible violations, and see also Schochet, 2001, for further discussion of them in Job Corps). By virtue of the exclusion restriction on the employment status, S, we can eliminate the $n\&EN$ group and the $n\&NE$ group, which would imply an effect of Z on S for noncompliers. The eight principal strata thus reduce to six: $c\&EE$, $c\&EN$, $c\&NE$, $c\&NN$, $n\&EE$, $n\&NN$.

Causal estimands of interest are usually, but not always, summaries of individual causal effects on a common set of units. Here we focus on the following average treatment effects in the super-population, because participants are randomly drawn from the population of eligible applicants, and the study has been conducted in order to inform policy makers on the effects of the program on such target super-populations. Specifically our estimands are:

- the average treatment effect of Z on program participation, D:

$$\Delta^{(ZD)} = E[D_i(1)|\boldsymbol{\theta}] = Pr[D_i(1) = 1|\boldsymbol{\theta}],$$

  which equals the proportion of compliers in the super-population;

- the average treatment effect of Z on employment, S:

$$\Delta^{(ZS)} = E[S_i(1)|\boldsymbol{\theta}] - E[S_i(0)|\boldsymbol{\theta}],$$

  which, by the exclusion restrictions, equals the difference of the proportions of EN and NE compliers in the super-population

$$Pr[G_i = c\&EN; \boldsymbol{\theta}] - Pr[G_i = c\&NE; \boldsymbol{\theta}];$$

- the average treatment effect of Z on employment, S, for compliers, which is usually interpreted as the effect of participation, D, on S:

$$\Delta^{(DS)} = E[S_i(1)|D_i(1) = 1; \boldsymbol{\theta}] - E[S_i(0)|D_i(1) = 1; \boldsymbol{\theta}]$$

$$= Pr[G_i = c\&EN|c; \boldsymbol{\theta}] - Pr[G_i = c\&NE|c; \boldsymbol{\theta}].$$

  In our analysis of JC, the effect of assignment for compliers is interpreted as the effect of immediate participation in JC relative to non-immediate participation, which may include no participation in any training program, participation in other training programs available on the market, or later participation in JC.

- the average treatment effect of Z on wages, W, for the always employed compliers, which is interpreted as the effect of participation on wages for the always-employed:

$$\Delta^{(DW)} = E[W_i(1)|G_i = c\&EE; \boldsymbol{\theta}] - E[W_i(0)|G_i = c\&EE; \boldsymbol{\theta}].$$

In the last two formulas, the expectations are taken over a subset of the entire super-population, the compliers for $\Delta^{(DS)}$ and the always-employed compliers for $\Delta^{(DW)}$.

The relative sizes in the population of the six principal strata are themselves relevant descriptive estimands:

$$Pr[G_i = g|\boldsymbol{\theta}], \quad g \in \mathcal{G}.$$

All previous estimands can be defined also conditional on specific values of some of the covariates. Policy-relevant information can also be obtained from our likelihood analysis about the distribution of baseline characteristics within each principal stratum, for example, the means of the covariates within strata:

$$\mu_{X,g} = E[X_i|G_i = g; \boldsymbol{\theta}], \quad g \in \mathcal{G}.$$

The ability to characterize the latent subgroups of units in terms of their initial conditions is an advantage of the approach we adopt, and may be particularly useful for targeting future interventions.

4 Mode of inference

4.1 Observed groups of units

Inference can be viewed as a missing data problem, because we cannot observe which stratum each unit belongs to. For each unit i, the treatment assignment indicator Z_i and the covariates $\mathbf{X}_i$ are always observed. Further, for unit i the other observed values are $M_i(Z_i) = M_{i,obs}$; if $M_i(Z_i) = 0$, $S_i(Z_i) = S_{i,obs}$ and $W_i(Z_i) = W_{i,obs}$; and for 99% of the units, we also observe $D_i(1)$ when $Z_i = 1$. We denote by $\mathbf{D(1)}$, $\mathbf{M_{obs}}$, $\mathbf{S_{obs}}$ and $\mathbf{W_{obs}}$ the corresponding vectors of observed values; note that if $M_{i,obs} = 1$ then $S_{i,obs} =?$ and $W_{i,obs} =?$; if $S_{i,obs} = 0$ then $W_{i,obs} = *$; if $Z_1 = 0$ then $D_i(1) =?$; if $Z_i = 1$ and $D_i(1)$ is missing, then $D_i(1) =?$.

Among units with observed outcomes, we can observe the following groups, defined according to different combinations of observed Z, D and S:

- $O(1,1,1) = \{i : Z_i = 1, D_i(1) = 1 \text{ and } S_{i,obs} = 1\}$, those who are assigned to the treatment group, take the treatment, and are employed; they are a mixture of the two principal strata $c\&EE$ and $c\&EN$;

- $O(1,1,0) = \{i : Z_i = 1, D_i(1) = 1 \text{ and } S_{i,obs} = 0\}$, those who are assigned to the treatment group, take the treatment, and are nonemployed; they are a mixture of the two principal strata $c\&NE$ and $c\&NN$;

- $O(1,0,1) = \{i : Z_i = 1, D_i(1) = 0 \text{ and } S_{i,obs} = 1\}$, those who are assigned to the treatment group, do not comply with assignment, and are employed; they belong to the principal stratum $n\&EE$;

- $O(1,0,0) = \{i : Z_i = 1, D_i(1) = 0 \text{ and } S_{i,obs} = 0\}$, those who are assigned to the treatment group, do not comply with assignment, and are not employed; they belong to the principal stratum $n\&NN$;

- $O(1,?,1) = \{i : Z_i = 1, D_i(1) =? \text{ and } S_{i,obs} = 1\}$, those who are assigned to the treatment group, with missing compliance status, and are employed; they are a mixture of the three

principal strata $c\&EE$, $c\&EN$, $n\&EE$;

- $O(1,?,0) = \{i : Z_i = 1, D_i(1) =? \text{ and } S_{i,obs} = 0\}$, those who are assigned to the treatment group, with missing compliance status, and are not employed; they belong to the three principal strata $c\&NE$, $c\&NN$, $n\&NN$;

- $O(0,?,1) = \{i : Z_i = 0, D_i(1) =? \text{ and } S_{i,obs} = 1\}$, those who are assigned to the control group and are employed; they are a mixture of the three principal strata $c\&EE, c\&NE, n\&EE$;

- $O(0,?,0) = \{i : Z_i = 0, D_i(1) =? \text{ and } S_{i,obs} = 0\}$, those who are assigned to the control group and are not employed; they are a mixture of the three principal strata $c\&EN, c\&NN, n\&NN$.

For units with missing outcomes, the values of S and W are not observed; the observed groups are:

- $O(1,1,?) = \{i : Z_i = 1, D_i(1) = 1 \text{ and } S_{i,obs} =?\}$, those who are assigned to the treatment group and take the treatment; they are a mixture of the four principal strata $c\&EE$, $c\&EN$, $c\&NE$, $c\&NN$;

- $O(1,0,?) = \{i : Z_i = 1, D_i(1) = 0 \text{ and } S_{i,obs} =?\}$, those who are assigned to the treatment group and do not comply with assignment; they are a mixture of the two principal strata $n\&EE, n\&NN$;

- $O(0,?,?) = \{i : Z_i = 0, D_i(1) =? \text{ and } S_{i,obs} =?\}$, those who are assigned to the control group and therefore have unknown compliance status; they are a mixture of all the principal strata in $\mathcal{G}$.

In Table 3, the correspondence between observed and latent groups is summarized.

4.2 Distributions for groups and potential outcomes given covariates

In order to form the likelihood function, we need to specify, first, a model for the principal stratum membership, G, given $\mathbf{X}$ and, second, the distribution of the potential wages conditional on G and $\mathbf{X}$. Note that $\mathbf{X}$ includes all the covariates, as defined in Table 1. For the covariates with a proportion

of missing data larger than 20% (parents' education and household income), we also included the missing indicators as well as interaction terms of those indicators with the covariates. The missing indicators may capture some salient features of the subjects: e.g., ignorance of parents' education may be associated with loss, absence of the parents' influence or extremely low parent's education, and thus may indicate a more disadvantaged situation. For the same reasons, missing indicators of these three variables were also used as fully observed covariates in the initial multiple imputation of missing covariates' values. Following Zhang *et al.* (2009), design weights are also included as a covariate in the models. By integrating the complete-data likelihood over the missing potential outcomes, under MAR, the observed data likelihood function is a finite mixture model likelihood (see, e.g., Imbens and Rubin, 1997), which can be maximized using the EM (Expectation-Maximization) algorithm (Dempster, Laird and Rubin, 1977). The steps of the EM algorithm used in the estimation procedure are derived and presented in the Appendix.

To simplify the notation, we assume that $\mathbf{X}$ includes the constant term – that is, a column containing the unit vector. We specify a multinomial logistic model for the k-dimensional vector of principal strata memberships:

$$Pr(G_i = g|\mathbf{X}_i;\boldsymbol{\alpha}) = \frac{\exp\{\mathbf{X}_i\boldsymbol{\alpha}_g\}}{\sum_{h=1}^{k}\exp\{\boldsymbol{X}_i\boldsymbol{\alpha}_h\}} = \pi_{i:g}$$

where $g \in G$, and the k^{th} principal stratum ($n\&NN$) is taken as the baseline (that is, $\boldsymbol{\alpha}_k = \mathbf{0}$), and let $\pi_{i:g}$ be the probability of belonging to stratum g for unit i, given the vector of pre-treatment covariates $\mathbf{X}_i$. Alternative specifications, such as sequential logistic models, multinomial probit models, or their t-based extensions (Liu, 2004), could also be used.

We specify a Normal distribution for log-wages conditional on covariates $\mathbf{X}$:

$$\begin{aligned}
&if\ G_i = c\&EE, && \log[W_i(1)] \sim N(\mathbf{X}_i\boldsymbol{\beta}_{c\&EE,1}, \sigma^2_{c\&EE,1}),\\
& && \log[W_i(0)] \sim N(\mathbf{X}_i\boldsymbol{\beta}_{c\&EE,0}, \sigma^2_{c\&EE,0}),\\
&if\ G_i = c\&EN, && \log[W_i(1)] \sim N(\mathbf{X}_i\boldsymbol{\beta}_{c\&EN,1}, \sigma^2_{c\&EN,1}),\\
&if\ G_i = c\&NE, && \log[W_i(0)] \sim N(\mathbf{X}_i\boldsymbol{\beta}_{c\&NE,0}, \sigma^2_{c\&NE,0}),\\
&if\ G_i = n\&EE, && \log[W_i(1)] \sim \log[W_i(0)] \sim N(\mathbf{X}_i\boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE}).
\end{aligned}$$

Notationally, we let $N_i(\mu, \sigma)$ be the probability density function of a Normal distribution with mean μ and variance σ evaluated at $\log(W_i)$. Zhang et al. (2009) already showed, for the same study, that alternative Box-Cox transformations (other than the logarithmic one and corresponding to different parametric families of distributions for wages) do not alter results significantly; so we maintained the assumptions of log-normality of wages conditional on principal strata, treatment assignment and a large set of covariates.

For the $c\&EE$ group, the parameters of the wage distribution vary between the treatment groups; for the $c\&EN$ group, wages are only defined on $\mathfrak{R}^+$ if $Z_i = 1$; for the $c\&NE$ group, wages are only defined on $\mathfrak{R}^+$ if $Z_i = 0$. The exclusion restriction implies that for the $n\&EE$ group, the distribution of wages is the same in the two treatment groups, therefore the parameters of their wage distributions are restricted to be the same irrespective of treatment assignment. For the $c\&NN$ and $n\&NN$ groups, there are no associated wage distributions defined on $\mathfrak{R}^+$ (i.e., they are $*$).

We denote by $\boldsymbol{\theta}_{sci} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}\}$ the vector parameter of scientific interest, a function of $\boldsymbol{\theta}$ assumed to be distinct from the function of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_{mis}$, governing the missingness mechanism, where

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{c\&EE}, \boldsymbol{\alpha}_{c\&EN}, \boldsymbol{\alpha}_{c\&NE}, \boldsymbol{\alpha}_{c\&NN}, \boldsymbol{\alpha}_{n\&EE}),$$

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_{c\&EE,1}, \boldsymbol{\beta}_{c\&EE,0}, \boldsymbol{\beta}_{c\&EN,1}, \boldsymbol{\beta}_{c\&NE,0}, \boldsymbol{\beta}_{n\&EE}),$$

$$\boldsymbol{\sigma} = (\sigma_{c\&EE,1}, \sigma_{c\&EE,0}, \sigma_{c\&EN,1}, \sigma_{c\&NE,0}, \sigma_{n\&EE}).$$

Assuming MAR, the observed data likelihood function is proportional to the joint distribution of $(\mathbf{D}(\mathbf{1}), \mathbf{S}_{obs}, \mathbf{W}_{obs} | \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta}_{sci})$:

$$L(\boldsymbol{\theta}_{sci} | \mathbf{D}(\mathbf{1}), \mathbf{S}_{obs}, \mathbf{W}_{obs}, \mathbf{Z}, \mathbf{X}) \propto$$

$$\prod_{i \in O(1,1,1)} \left[\pi_{i:c\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&EE,1}, \sigma^2_{c\&EE,1}) + \pi_{i:c\&EN} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&EN,1}, \sigma^2_{c\&EN,1}) \right]$$

$$\times \prod_{i \in O(1,1,0)} [\pi_{i:c\&NE} + \pi_{i:c\&NN}] \times \prod_{i \in O(1,0,1)} \left[\pi_{i:n\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE}) \right] \times \prod_{i \in O(1,0,0)} [\pi_{i:n\&NN}]$$

$$\times \prod_{i \in O(1,?,1)} \left[\pi_{i:c\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&EE,1}, \sigma^2_{c\&EE,1}) + \pi_{i:c\&EN} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&EN,1}, \sigma^2_{c\&EN,1}) + \pi_{i:n\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{n\&EE,1}, \sigma^2_{n\&EE,1}) \right]$$

$$\times \prod_{i \in O(1,?,0)} [\pi_{i:c\&NE} + \pi_{i:c\&NN} + \pi_{i:n\&NN}]$$

$$\times \prod_{i \in O(0,?,1)} \left[\pi_{i:c\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&EE,0}, \sigma^2_{c\&EE,0}) + \pi_{i:c\&NE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&NE,0}, \sigma^2_{c\&NE,0}) + \pi_{i:n\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE})\right]$$

$$\times \prod_{i \in O(0,?,0)} [\pi_{i:c\&EN} + \pi_{i:c\&NN} + \pi_{i:n\&NN}]$$

$$\times \prod_{i \in O(1,1,?)} [\pi_{i:c\&EE} + \pi_{i:c\&EN} + \pi_{i:c\&NE} + \pi_{i:c\&NN}] \times \prod_{i \in O(1,0,?)} [\pi_{i:n\&EE} + \pi_{i:n\&NN}] .$$

The units in the $O(0, ?, ?)$ group do not carry any information and vanish from the likelihood function (because $\sum_g \pi_{i:g} = 1$). The likelihood function of normal mixture models is not a bounded function on the usual parameter space (Kiefer and Wolfowitz, 1956; Day, 1969). However, in spite of this unboundedness, Peters and Walker (1978) prove that, given any sufficiently small neighborhood of the true parameter, with probability one, the MLE exists, it is unique and it is (locally) strongly consistent. Redner (1981) proves that the MLE exists and it is globally consistent in every compact parameter subset containing the true value of the parameter. Consequently, we can exploit standard mixture model analysis (e.g., see Titterington *et al.*, 1985) for identification and inference. Except in the very special case when the proportions of the mixture components are the same, we can uniquely estimate the mixture parameters (see also Everitt and Hand, 1981, Gelman *et al.*, 2004, and Zhang *et al.*, 2009).

As a robustness check, alternative nonignorable missingness assumptions are considered in the empirical analysis, based on different behavioral hypotheses on the missingness mechanism. Details on these assumptions and the ML analysis under them are presented in the Appendix.

4.3 Estimation of causal estimands

Causal effects are estimated in each principal stratum as functions of the observed data and the ML estimates of parameters, averaging over the estimated population distribution of covariates in that principal stratum using the design weights to weight the N units in the sample to represent the superpopulation. More explicitly, letting $\hat{\pi}_{i:g} = Pr(G_i = g|\boldsymbol{X}_i, \hat{\boldsymbol{\alpha}})$, we estimate the proportion in each stratum as

$$\hat{\pi}_g = \frac{\sum_{i=1}^N \omega_i \hat{\pi}_{i:g}}{\sum_{i=1}^N \omega_i},$$

where ω_i are design weights in the original survey. Then, the causal effect of Z on D is estimated as the proportion of compliers: $\hat{\Delta}^{(ZD)} = \hat{\pi}_{c\&EE} + \hat{\pi}_{c\&EN} + \hat{\pi}_{c\&NE} + \hat{\pi}_{c\&NN} = \hat{\pi}_c$, and the causal effect of Z on S is estimated as the difference of the estimated proportions of EN and NE compliers: $\hat{\Delta}^{(ZS)} = \hat{\pi}_{c\&EN} - \hat{\pi}_{c\&NE}$. Similarly, estimates of the average treatment effects on employment for compliers, and on wages for always-employed compliers are obtained as:

$$\hat{\Delta}^{(DS)} = \frac{\hat{\pi}_{c\&EN} - \hat{\pi}_{c\&NE}}{\hat{\pi}_c}$$

and

$$\hat{\Delta}^{(DW)} = \frac{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:c\&EE} \exp\left\{\mathbf{X}_i \hat{\boldsymbol{\beta}}_{c\&EE,1} + \frac{1}{2}\hat{\sigma}^2_{c\&EE,1}\right\}}{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:c\&EE}} - \frac{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:c\&EE} \exp\left\{\mathbf{X}_i \hat{\boldsymbol{\beta}}_{c\&EE,0} + \frac{1}{2}\hat{\sigma}^2_{c\&EE,0}\right\}}{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:c\&EE}}$$

respectively.

In order to characterize the latent subgroups, the means of the covariates within each principal stratum are estimated as follows:

$$\hat{\mu}_{X,g} = \frac{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:g} X_i}{\sum_{i=1}^{N} \omega_i \hat{\pi}_{i:g}}.$$

By using the estimated principal strata membership proportion $\hat{\pi}_{i:g}$, we are implicitly imputing the potential outcome for given ω_i and $\boldsymbol{X}_i$ an infinite number of times, thus estimating superpopulation parameters. If the asymptotic covariance matrix of the estimates were obtained, the asymptotic standard errors of the above quantities could be computed using the Delta method or methods such as the SEM algorithm (Meng and Rubin, 1991). However, even in relatively large samples, the sampling distribution of ML estimators is usually not well approximated by the standard asymptotic normal distribution because the likelihood function for mixture models is generally not close to normal. For this reason, we focussed on comparisons of the maximized likelihood function under the general model and under various meaningful restrictions using a direct likelihood approach advocated in some situations by Fisher (1921), Barnard *et. al* (1962), Barnard (1965), Hacking (1965), Edwards (1972), Royall (1997), and recently discussed in Boyles (2008) and used in Zhang et al. (2009) in a related problem.

5 Likelihood-based estimation: results

5.1 Likelihood estimation of causal estimands under MAR

We discuss results obtained by maximizing the likelihood function and provide ML estimates of the average treatment effects on employment ($\Delta^{(DS)}$) and wages ($\Delta^{(DW)}$) for each week separately, together with the estimated proportions of the principal strata, as described in Section 4.3. We do not report maximum likelihood estimates of the model parameters, which are however available upon request from the authors. Most of the covariates' coefficients have the expected sign; for example, higher educated people tend to have higher wages irrespective of their principal stratum. We also compare the maximized likelihood under the general model with the maximized likelihoods under three meaningful restrictions: (a) monotonicity of employment: $\pi_{c\&NE} = 0$; (b) no effect of assignment on employment for compliers: $\Delta^{(DS)} = 0$; and (c) no effect of assignment on wages for the always-employed compliers: $\Delta^{(DW)} = 0$. Specifically, Table 4 presents values of the scaled log-likelihood ratio statistic, λ, for the general model versus models with restrictions, calculated as $-2log(\Lambda)/df$, where Λ is the ratio of the maximized likelihood under the general model and under a model with specific restrictions, and where df is equal to the difference of the number of parameters in the models. A strong deviation of this quantity from one provides evidence that the corresponding restriction is not supported by the data.

The overall results on average causal estimands suggest the following summaries.

First, monotonicity of employment is not supported by the data at any week (see the values of λ_M in Table 4), suggesting that all assumed six latent strata exist, and that there is a positive proportion of compliers, $\pi_{c\&NE}$, for whom training appears detrimental in terms of employment. A possible conjecture is that these people might have raised their reservation wages as a consequence of training, and refuse job offers that would be accepted with no training. As expected, the proportions of $c\&NE$ decreases over time, and the non-negligible percentage of them four years after assignment may be simply due to the structural mobility in and out of employment of American youths.

Second, the proportions of $\pi_{c\&NE}$ and $\pi_{c\&EN}$ appear to be roughly equal to each other at all three

time points, thus suggesting that the average effect of assignment on employment for compliers is absent. In fact, the restriction of no average effect of assignment on employment appears to be plausible at all three weeks (see the values of λ_{0W} in Table 4). However, the point estimates of these effects are larger than the ITT effects on employment found in Zhang et al. (2009) at week 208, which were diluted by noncompliance to treatment assignment. The negative point estimate of the effect on employment in the short term (of about -2.4% at week 52) and the positive ones in the long term (of about 2.2% at week 130 and 1.3% at week 208) are consistent with the empirical literature on the effect of active labor market policies, which suggests that almost all programs reduce employment in the short run (e.g., Lechner and Wunsch, 2007; van Ours, 2004). Note that, by looking at the sizes of the two groups of $c\&NE$ and $c\&EN$, instead of at the overall effect on employment for compliers, our analysis offers a more refined understanding of how such a small estimated effect on employment was produced.

Third, no effect of assignment on wages for the always-employed compliers is rejected by the data at all weeks (see values of λ_{S0} in Table 4): From Table 5, we see that the average effect of assignment on wages is found to be small but positive (about 0.28, 0.25 and 0.29 \$/hour at weeks 52, 130 and 208, respectively), corresponding to approximately 4 to 5% increases relative to the average wage with no JC. Again, this is a different finding from Zhang et al. (2009), where the effect of assignment on wages for the always-employed at week 208 was found to be negligible, after discarding units with missing outcomes, thus not adhering to the ITT principle. Note that, although the effect on employment is ideally estimated for the same group of units over the three weeks, i.e., compliers, the effects on wages are for the latent group of the always-employed compliers, which includes different units at different weeks.

Our results deviate from the naive conclusions one could draw from simple descriptive contrasts presented in Section 2; naive comparisons and simple IV comparisons, which neglect some of the complications, appear to overestimate the impact of the program. Differences between these contrasts and the estimated causal effects are larger at week 208, especially for wages, possibly due to the larger missingness rate observed at this week.

These are general overall results, which offer information on the effects of assignment for

compliers, and thus the effects of participation in the program. However, simply looking at average effects limits the usefulness of the results, which do not offer particularly strong evidence in favor of the effectiveness of Job Corps, yet do not provide any constructive information to help understand what could be improved in the implementation of such a program. The framework we adopted, however, is not only a proper one for formally dealing with the complications of JC, but it also allows one to exploit the presence of these complications to extract additional information from the data. Specifically, further insights into the principal strata can be obtained by analyzing both the distribution of background characteristics and the distribution of wages within the strata; those analyses can generate useful suggestions for the re-design of the program. In Table 5, the estimated average wages for all strata under treatment and under control are reported, along with asymptotic standard errors, in order to have a rough quantification of the sampling variability. In Tables 6, 7 and 8, the estimated means of the covariates within each stratum are reported, obtained using, for each unit, the design weights and the estimated membership probabilities.

The distribution of covariates among noncompliers suggests that the reasons for noncompliance may differ, implying that better suited programs should have been offered to different subjects. The average characteristics of the $n\&EE$ individuals show that they are in general older and better educated with longer labour market experience: most of them already worked, had longer tenure in previous jobs and were better paid. They thus appear to be people, on average, who should not have been targets of the program in the first place. Conversely, the never-employed noncompliers, $n\&NN$, are in general less likely to be white and more likely to be female and have children; they appear to be the right target of the program, and so their decision to not participate in the program may be partly explained by objective difficulties of participation due to family constraints, suggesting a more flexible training schedule for them may have satisfied their requirements.

Regarding the groups who participated in the program, the never-employed compliers, $c\&NN$, are, in general, less likely to be well educated or white, they had shorter tenure in previous jobs and were paid less. They appear to be mostly disadvantaged individuals, with the worst average initial conditions. For them, participation in the program was not beneficial in terms of employment, suggesting a re-designed intervention for them, even more focussed on the disadvantaged

participants, allowing them to improve their educational levels and acquire job-specific skills, or providing help with their job-search activities.

Compliers who are "sometimes employed" ($c\&NE$ and $c\&EN$) are generally less likely to be female, to have children, or to have a partner; and more likely to be white. These characteristics suggest that these subjects are more mobile in the labor market and have fewer constraints than others; they are thus more likely to be observed without a job. This finding is also consistent with the evidence from Table 5 that these compliers have higher post JC average wages than the always employed compliers, $c\&EE$, suggesting that these groups comprise individuals who are more selective when deciding whether to accept a job offer: they tend to have better paid but less stable jobs. This possibly mitigates the apparently disappointing result of a small effect of participation on employment.

The group of the always-employed compliers, $c\&EE$, for whom the effect on wages was sought, does not show striking differences from the other compliers, except for the level of their wages, which from Table 5 appears to be lower, under either treatment or control, than the other groups when employed, $c\&NE$, $c\&EN$, and $n\&EE$. The effect on wages for them is a positive and stable one, so that, for the $c\&EE$ subgroup, the program was mildly successful, in absolute and relative terms, in increasing labor productivity reflected in a wage increase.

In order to assess the robustness of our results to deviations from ignorability, we also provide estimates of causal effects, as well as the values of the scaled LRT statistics, obtained by maximizing the likelihood under Latent Ignorability (LI) and two different sets of restrictions, as detailed in the Appendix. Results, reported in Table 9, show that the estimates of the relevant treatment effects are not sensitive to these three alternative assumptions on the missing data mechanism. In addition, the values of the scaled LRT statistics show that the data, also under LI, neither support monotonicity of employment nor a null effect on wages for the always employed compliers, but the data do support a null effect on employment for compliers. We argue that this substantial similarity of results, under ignorable and simple nonignorable models, is because we always condition on a rich set of baseline characteristics, which provide for relatively good predictions of the latent principal strata and of the missing potential outcomes. They also mitigate distributional

assumptions on the potential outcomes within strata. Once we condition on these covariates, the missingness assumptions (as well as the distributional assumptions, as highlighted in Zhang *at. al*, 2009) make only minor differences, which stresses the importance of collecting baseline characteristics in experimental studies: they may help deal with subsequent complications that "break" the initial randomization.

5.2 Discussion

The framework we used, and the tools we developed, are appropriate for conducting an even more comprehensive longitudinal analysis. This, however, would have implied a growing number of principal strata, so we analyzed the three weeks separately. Even so, consistent results were obtained. For example, the percentage of noncompliers, which should be constant across all weeks, but not constrained to be so in our analysis, is estimated from Table 3 to be around $28-29\%$ at all weeks. This result can be seen as a simple diagnostic for the fit of our models.

Our analysis not only allowed the assessment of the overall effects of the program, but also the assessment of whether the program was well targeted, for whom the program worked best, and for which outcome. In fact, a policy relevant result obtained in this paper was the ability to characterize the latent subgroups in terms of their initial pre-treatment conditions. The most disadvantaged groups, with the worst average initial conditions in terms of education, labour market experience, race and gender, are the never-employed ($c\&NN$ and $n\&NN$), who did not benefit from the training program even when they decided to participate ($c\&NN$). The groups of compliers who benefited from participation in terms of employment or wages appear to be less disadvantaged on average than the never-employed.

These findings may be useful to help re-design the program for better effectiveness: the nuanced results resolve much of the interpretational issues because they directly inform the policy-maker about whether the program was well-targeted, whether it was uniformly effective for all the subjects, and about which of its objectives may have been achieved. In fact, as with most of the large job-training programs, Job Corps had, and still has, different aims and employed a mixture of instruments to try to reach them. Training activities may be specifically targeted at particular

groups (e.g., the young or the disabled), may be designed to prevent long periods out of regular employment, or to integrate unemployed and disadvantaged individuals into the labour force, or they may be more oriented towards augmenting participants' human capital, either by helping them earn a higher educational degree or by formal teaching of new vocational skills. Job Corps employs "a holistic career development training approach which [sic] integrates the teaching of academic, vocational, employability skills and social competencies through a combination of classroom, practical and based learning experiences to prepare youth for stable, long-term, high-paying jobs". From our findings, Job Corps seems to have been successful only in augmenting participants' human capital, as measured by the effects on wages for the always-employed compliers, although it does not seem to have enhanced the employability of the more disadvantaged.

References

[1] Angrist, Joshua D. (1990). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administration Records. American Economic Review, 80, 313-336.

[2] Angrist, J.D., Imbens, G.W., Rubin, D.B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91, 444-455.

[3] Barnard, G. A. (1965). The Use of the Likelihood Function in Statistical Practice. Proceedings of the Fifth Berkeley Symposium, 1, 27-40.

[4] Barnard, G. A., Jenkins, G. M., and Winsten, C. B. (1962). Likelihood Inference and Time Series. Journal of the Royal Statistical Society, Ser. A, 125, 321372.

[5] Boyles, R. A. (2008). The Role of Likelihood in Interval Estimation. The American Statistician, 62, 2226.

[6] Burghardt, J., McConnell, S., Schochet, P., Johnson, T., Gritz, M., Glazerman, S., Homrighausen, J. (2001). Does Job Corps work? Summary of the National Job Corps Study. Document No. PR01-50, Princeton, NJ: Mathematica Policy Research, Inc.

[7] Day, N. E. (1969). Estimating the components of a mixture of normal distributions, Biometrika 56, 463-474.

[8] Deaton, A. (2010). Instruments, Randomization, and Learning about Development. Journal of Economic Literature, 48, 424455.

[9] Dehejia, R. H., Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. Journal of the American Statistical Association, 94, 105362.

[10] Dempster, A.P., Laird, N., Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data using the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B 39, 1-38.

[11] Edwards, A. W. F. (1972). Likelihood. Cambridge, UK: Cambridge University Press.

[12] Everitt B. S. and Hand D. J. (1981). Finite Mixture Distributions. London: Chapman and Hall.

[13] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). Bayesian Data Analysis. Second edition. London: Chapman and Hall.

[14] Fisher, R. A. (1921). On the Probable Error of a Coefficient of Correlation Deduced From a Small Sample. Metron, 1, 332.

[15] Flores C.A., Flores-Lagunes A. (2009). Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness. IZA Discussion Paper, No. 4237.

[16] Flores-Lagunes, A., Gonzalez, A., Neumann, T. (2007). Estimating the effects of length of exposure to a training program: the case of Job Corps. IZA Discussion Paper, No. 2846.

[17] Frangakis, C.E., Rubin, D.B. (2002). Principal stratification in causal inference. Biometrics, 58, 21-29.

[18] Hacking, I. (1965). Logic of Statistical Inference. Cambridge University Press, New York.

[19] Heckman, J.J. (1979). Sample selection bias as a specification error. Econometrica, 47,153-162.

[20] Heckman, J.J., Lalonde, R.J., Smith, J.A. (1999). The economics and econometrics of active labor market programs. In: Handbook of Labor Economics, Vol.3, Ashenfelter, O., Card, D. eds.

[21] Heckman, J.J. (2010). Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy Journal of Economic Literature, 48, 356-398.

[22] Holland, P. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 945-960.

[23] Imbens, G.W. (2010). Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). Journal of Economic Literature, 48.

[24] Imbens, G.W., Angrist, J. (1994). Identification and estimation of local average treatment effects. Econometrica, 62, 467-476.

[25] Imbens, G.W., Rubin, D.B. (1997). Estimating outcome distributions for compliers in instrumental variables models. Review of Economic Studies, 64, 555-574.

[26] Jin, H., Rubin, D.B. (2009). Public Schools versus Private Schools: Causal Inference with Partial Compliance. Journal of Educational and Behavioral Statistics, 34, 24-45.

[27] Kiefer, J. and Wolfowitz, J. (1956). Consistency on the maximum likelihood estimator in the presence of Infinitly many incidental parameters, The Annals of Mathematical Statistics 27, 887-906.

[28] Lechner, M., Wunsch, C. (2007). Are training programs more effective when unemployment is high? IAB Discussion Paper 200707.

[29] Lee, D.S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. Review of Economic Studies, 76, 1071-1102.

[30] Little, R. J. A. and Rubin, B. D. (2002). Statistical Analysis with Missing Data, 2nd ed. Wiley, New York.

[31] Liu, C. (2004). Robit Regression: A Robust Alternative to Logit and Probit. In Missing Data and Bayesian Methods in Practice: Contributions by Donald Rubins Statistical Family, eds. A. Gelman and X. L. Meng, Wiley, New York.

[32] Liu, C., and Rubin, D. B. (1998). Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data. Biometrika, 85, 673688.

[33] McConnell, S., Stuart, E.A., Devaney, B.m. (2008). The Truncation-by-Death Problem. What To Do in an Experimental Evaluation When the Outcome Is Not Always Defined. Evaluation Review, 32, 157-186.

[34] Mealli F., Imbens G., Ferro S., and Biggeri A. (2004). Analyzing a Randomized Trial on Breast Self Examination with Noncompliance and Missing Outcomes, Biostatistics, 5(2), 207–222.

[35] Meng, Rubin, D.B. (1991). Using EM to obtain variance-covariance matrices: the SEM algorithm.Journal of teh American Statistica Association, 86, 899-909.

[36] Peters, B.C. and Walker, H.F. (1978). An iterative procedure for obtaining maximum likelihood estimation of the parameters for a mixture of normal distributions, SIAM Journal of Applied Mathematics 35, 362-378.

[37] Redner, R. (1981). Note on the consistency of the maximum like lihood estimate for nonidentifiable distributions, The Annals of Statistics 9, 225-228.

[38] Royall, R. M. (1997), Statistical Evidence, Chapman and Hall, London.

[39] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non randomized studies. Journal of Educational Psychology, 66, 688-701.

[40] Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581-592.

[41] Rubin, D.B. (1978). Bayesian inference for causal effects. Annals of Statistics, 6, 34-58.

[42] Rubin, D.B. (1980). Discussion of "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu. Journal of the American Statistical Association, 75, 591-593.

[43] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys, Wiley, New York.

[44] Rubin, D.B. (1990). Formal modes of statistical inference for causal effects. Journal of Statistical Planning and Inference, 25, 279-292.

[45] Rubin, D.B. (2000). The utility of counterfactuals for causal Inference – Discussion of "Causal inference without counterfactuals" by A.P. Dawid. Journal of the American Statistical Association, 95, 435-438.

[46] Rubin, D.B. (2005). Causal Inference using potential outcomes: design, modeling, decisions. Journal of the American Statistical Association, 100, 322-331.

[47] Rubin, D.B. (2006). Causal inference through potential outcomes and principal stratification: application to studies with censoring due to death. Statistical Science, 21, 299-321.

[48] Schochet, P. Z. (2001). National Job Coprs Study: Methodological Appendixes on the Impact Analysis. Mathematica Policy Research, Inc.

[49] Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). Statistical Analysis of Finite Mixture Distributions, New York: Wiley.

[50] van Buuren, S., Oudshoorn C. G. M. (1999). Flexible multivariate imputation by MICE, Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054 .

[51] van Ours, J. (2004). The locking-in effect of subsidized jobs. Journal of Comparative Economics, 32, 37-52.

[52] Zhang, J.L., Rubin, D.B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". Journal of Educational and Behavioral Statistics, 28, 353-368.

[53] Zhang, J.L., Rubin, D.B., Mealli, F. (2008). Evaluating the effects of job training programs on wages through principal stratification. In Modelling and Evaluating Treatment Effects in Econometrics, D.L. Millimet, J.A. Smith, E.J. Vytlacil, eds., Elsevier, 117-145.

[54] Zhang, J.L., Rubin, D.B., Mealli, F. (2009). Likelihood-based analysis of the causal effects of job-training programs using principal stratification. Journal of the American Statistical Association, 104, 166-176.

Appendix

Ignorable and nonignorable missing data mechanisms.

The joint distribution of observable potential outcomes given $\mathbf{X}$, $f(\mathbf{D}(1), \mathbf{S}(0), \mathbf{S}(1), \mathbf{W}(0), \mathbf{W}(1), \mathbf{M}(0), \mathbf{M}(1)|\mathbf{X})$ (see equation 2), can be decomposed into one factor modelling the quantities of scientific interest

$$f_{sci}(\mathbf{D}(1), \mathbf{S}(0), \mathbf{S}(1), \mathbf{W}(0), \mathbf{W}(1)|\mathbf{X}, \boldsymbol{\theta}_{sci}),$$

and one factor representing the missingness mechanism, i.e., the distribution of the missing indicators given the other potential outcomes and covariates: $f_{mis}(\mathbf{M}(0), \mathbf{M}(1)|\mathbf{D}(1), \mathbf{S}(0), \mathbf{S}(1), \mathbf{W}(0), \mathbf{W}(1), \mathbf{X}, \boldsymbol{\theta}_{mis})$, where $\boldsymbol{\theta}_{sci}$ and $\boldsymbol{\theta}_{mis}$ are the functions of $\boldsymbol{\theta}$ governing the corresponding distributions. Under MAR and if the parameters of the missing data mechanism, $\boldsymbol{\theta}_{mis}$), are distinct from those of the outcome distributions, $\boldsymbol{\theta}_{sci}$), the missing data process is ignorable (Rubin, 1976), meaning that valid likelihood inference ignores the missing data model.

Nonignorable missing-data mechanisms are often difficult to specify because there is rarely direct evidence in the data about the relationship between the missing-data mechanism and the missing values themselves. It is usually advisable to consider several nonignorable models, and to explore the sensitivity of estimates of relevant causal estimands to the different models, using a baseline analysis under MAR as a primary benchmark for comparison. In a PS framework, a plausible nonignorable missingness assumption is "Latent Ignorability" (LI), originally proposed by Frangakis and Rubin (1999) in a setting with noncompliance. Because here the scientifically relevant principal strata, G_i, are defined according to noncompliance and potential employment status, we formulate Latent Ignorability to mean that, if we knew the group membership (G_i) of each unit, the missingness mechanism would be ignorable:

$$f_{mis}(M_i(0), M_i(1)|D_i(1), S_i(0), S_i(1), W_i(0), W_i(1), \mathbf{X}_i, \boldsymbol{\theta}_{mis}) = \tag{3}$$

$$f_{mis}(M_i(0), M_i(1)|G_i, W_i(0), W_i(1), \mathbf{X}_i, \boldsymbol{\theta}_{mis}) = f_{mis}(M_i(0), M_i(1)|G_i, \mathbf{X}_i, \boldsymbol{\theta}_{mis}).$$

Under this assumption, given the covariates and treatment assignment, units with the same compliance behavior and potential employment status (and so with the same value of G_i) are expected to have the same distribution of wages, regardless of their missingness behavior. However, because the true compliance behaviors and the potential employment statuses are partially unobserved, the missing data process is nonignorable. We assume that the joint distribution of $M_i(0)$, $M_i(1)$, $f_{mis}(M_i(0), M_i(1)|G_i, \mathbf{X}_i, \boldsymbol{\theta}_{mis})$, has twelve independent parts, two for each of the six principal strata defined by G_i, which are assumed to be conditionally independent. Specifically, let

$$\rho_{i:g,z} = Pr(M_i(z) = 1|G_i = g, \mathbf{X}_i; \boldsymbol{\theta}_{mis}^{g,z})$$

be the probability of missing outcomes for unit i, $i = 1, ..., N$, when assigned treatment z, conditional on principal stratum membership g and $\mathbf{X}_i$; $\boldsymbol{\theta}_{mis} = \{\boldsymbol{\theta}_{mis}^{g,z}\}$, $(g \in \mathcal{G}, z = \{0, 1\})$. These probabilities can be regarded as nuisance unknowns of little intrinsic scientific interest and, depending on the empirical context, modified versions of exclusion restrictions for them found in the literature may be plausible.

The first exclusion restriction assumes that compliers have the same probability of having a missing outcome irrespective of treatment assignment and employment status; similarly noncompliers have the same same probability of having a missing outcome irrespective of employment status, but these probabilities are allowed to differ by treatment assignment. This is a similar assumption to the response exclusion restriction for compliers proposed in Mealli et al. (2004):

$$\begin{gathered}\rho_{i:c\&EE,1} = \rho_{i:c\&EN,1} = \rho_{i:c\&NE,1} = \rho_{i:c\&NN,1} = \rho_{i:c\&EE,0} = \rho_{i:c\&EN,0} = \rho_{i:c\&NE,0} = \rho_{i:c\&NN,0}, \\ \rho_{i:n\&EE,1} = \rho_{i:n\&NN,1} \\ \rho_{i:n\&EE,0} = \rho_{i:n\&NN,0}.\end{gathered} \tag{4}$$

Because compliers are willing to follow the protocol in their assigned treatment, it seems more plausible that the missingness mechanism would not be affected by that assignment and the subsequent employment status.

The second exclusion restriction posits that noncompliers have the same probability of having a missing outcome irrespective of their treatment assignment and post-treatment employment status; similarly compliers have the same probability of having a missing outcome irrespective of employment status, but these probabilities are allowed to differ in the two treatment arms. This is a similar assumption to the response exclusion restriction for never-takers in Frangakis and Rubin (1999):

$$\begin{gathered}\rho_{i:c\&EE,1} = \rho_{i:c\&EN,1} = \rho_{i:c\&NE,1} = \rho_{i:c\&NN,1} \\ \rho_{i:c\&EE,0} = \rho_{i:c\&EN,0} = \rho_{i:c\&NE,0} = \rho_{i:c\&NN,0} \\ \rho_{i:n\&EE,1} = \rho_{i:n\&NN,1} = \rho_{i:n\&EE,0} = \rho_{i:n\&NN,0}.\end{gathered} \tag{5}$$

Under LI and either of these two sets of Assumptions (4 or 5), the missingness mechanism is not ignorable because the missingness probabilities do not factor out of the likelihood.

Specifically, let $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\theta}_{mis}\}$ denote the vector parameter, where $\boldsymbol{\theta}_{mis}$ is the sub-vector of parameters governing the missingness probabilities in (3). Assuming LI, the likelihood function is derived as proportional to the joint distribution of $(\mathbf{D}(\mathbf{1}), \mathbf{M}_{obs}, \mathbf{S}_{obs}, \mathbf{W}_{obs} | \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta})$:

$$\begin{aligned}&L(\boldsymbol{\theta}|\mathbf{D}(\mathbf{1}), \mathbf{M}_{obs}, \mathbf{S}_{obs}, \mathbf{W}_{obs}, \mathbf{Z}, \mathbf{X}) \propto \\ &\prod_{i \in O(1,1,1)} \left[\bar{\rho}_{i:c\&EE,1}\pi_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&EE,1}, \sigma^2_{c\&EE,1}) + \bar{\rho}_{i:c\&EN,1}\pi_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&EN,1}, \sigma^2_{c\&EN,1})\right] \\ &\times \prod_{i \in O(1,1,0)} \left[\bar{\rho}_{i:c\&NE,1}\pi_{i:c\&NE} + \bar{\rho}_{i:c\&NN,1}\pi_{i:c\&NN}\right] \times \prod_{i \in O(1,0,1)} \left[\bar{\rho}_{i:n\&EE,1}\pi_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE})\right] \\ &\times \prod_{i \in O(1,0,0)} \left[\bar{\rho}_{i:n\&NN,1})\pi_{i:n\&NN}\right] \times \prod_{i \in O(1,?,1)} \left[\bar{\rho}_{i:c\&EE,1}\pi_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&EE,1}, \sigma^2_{c\&EE,1})\right. \\ &\left. + \bar{\rho}_{i:c\&EN,1}\pi_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&EN,1}, \sigma^2_{c\&EN,1}) + \bar{\rho}_{i:n\&EE,1}\pi_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE})\right] \\ &\times \prod_{i \in O(1,?,0)} \left[\bar{\rho}_{i:c\&NE,1}\pi_{i:c\&NE} + \bar{\rho}_{i:c\&NN,1}\pi_{i:c\&NN} + \bar{\rho}_{i:n\&NN,1}\pi_{i:n\&NN}\right]\end{aligned} \tag{6}$$

$$\times \prod_{i \in O(0,?,1)} \left[\bar{\rho}_{i:c\&EE,0} \pi_{i:c\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&EE,0}, \sigma^2_{c\&EE,0}) \right.$$

$$\left. + \bar{\rho}_{i:c\&NE,0} \pi_{i:c\&NE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&NE,0}, \sigma^2_{c\&NE,0}) + \bar{\rho}_{i:n\&EE,0} \pi_{i:n\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE}) \right]$$

$$\times \prod_{i \in O(0,?,0)} \left[\bar{\rho}_{i:c\&EN,0} \pi_{i:c\&EN} + \bar{\rho}_{i:c\&NN,0} \pi_{i:c\&NN} + \bar{\rho}_{i:n\&NN,0} \pi_{i:n\&NN} \right]$$

$$\times \prod_{i \in O(1,1,?)} \left[\rho_{i:c\&EE,1} \pi_{i:c\&EE} + \rho_{i:c\&EN,1} \pi_{i:c\&EN} + \rho_{i:c\&NE,1} \pi_{i:c\&NE} + \rho_{i:c\&NN,1} \pi_{i:c\&NN} \right]$$

$$\times \prod_{i \in O(1,0,?)} \left[\rho_{i:n\&EE,1} \pi_{i:n\&EE} + \rho_{i:n\&NN,1} \pi_{i:n\&NN} \right] \times \prod_{i \in O(0,?,?)} \left[\sum_{g \in \mathcal{G}} \rho_{i:g,0} \pi_{i:g} \right].$$

where $\bar{\rho}_{i:g,z} = 1 - \rho_{i:g,z}$, and the missingness probabilities are specified using binary logistic models:

$$\rho_{i:g,z} = \frac{\exp\{\mathbf{X}_i \boldsymbol{\theta}^{g,z}_{mis}\}}{1 + \exp\{\mathbf{X}_i \boldsymbol{\theta}^{g,z}_{mis}\}}.$$

Missingness probabilities are regarded as nuisance unknowns, in contrast to parameters of the outcome principal strata distributions, which define causal estimands of interest. We maximize the likelihood under each of the two sets of restrictions (4) and (5).

EM steps under MAR

The complete-data log-likelihood function given the principal strata G_i, i.e. treating G as the missing data, under ignorability can be written as follows:

$$l(\boldsymbol{\theta}_{sci} | \mathbf{D}(1), \mathbf{M}_{obs}, \mathbf{S}_{obs}, \mathbf{W}_{obs}, \mathbf{G}, \mathbf{Z}, \mathbf{X}) \propto$$

$$\sum_{i \in O(1,1,1)} I(G_i = c\&EE) \log \left[\pi_{i:c\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&EE,1}, \sigma^2_{c\&EE,1}) \right]$$

$$+ \sum_{i \in O(1,1,1)} I(G_i = c\&EN) \log \left[\pi_{i:c\&EN} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&EN,1}, \sigma^2_{c\&EN,1}) \right]$$

$$+ \sum_{i \in O(1,1,0)} I(G_i = c\&NE) \log \left[\pi_{i:c\&NE} \right] + \sum_{i \in O(1,1,0)} I(G_i = c\&NN) \log \left[\pi_{i:c\&NN} \right]$$

$$+ \sum_{i \in O(1,0,1)} I(G_i = n\&EE) \log \left[\pi_{i:n\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE}) \right]$$

$$+ \sum_{i \in O(1,0,0)} I(G_i = n\&NN) \log \left[\pi_{i:n\&NN} \right]$$

$$+ \sum_{i \in O(1,?,1)} I(G_i = c\&EE) \log \left[\pi_{i:c\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&EE,1}, \sigma^2_{c\&EE,1}) \right]$$

$$+ \sum_{i \in O(1,?,1)} I(G_i = c\&EN) \log \left[\pi_{i:c\&EN} N_i(\mathbf{X}_i \boldsymbol{\beta}_{c\&EN,1}, \sigma^2_{c\&EN,1}) \right]$$

$$+ \sum_{i \in O(1,?,1)} I(G_i = n\&EE) \log \left[\pi_{i:n\&EE} N_i(\mathbf{X}_i \boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE}) \right]$$

$$
\begin{aligned}
&+ \sum_{i \in O(1,?,0)} I(G_i = c\&NE) \log\left[\pi_{i:c\&NE}\right] \\
&+ \sum_{i \in O(1,?,0)} I(G_i = c\&NN) \log\left[\pi_{i:c\&NN}\right] \\
&+ \sum_{i \in O(1,?,0)} I(G_i = n\&NN) \log\left[\pi_{i:n\&NN}\right] \\
&+ \sum_{i \in O(0,?,1)} I(G_i = c\&EE) \log\left[\pi_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&EE,0}, \sigma^2_{c\&EE,0})\right] \\
&+ \sum_{i \in O(0,?,1)} I(G_i = c\&NE) \log\left[\pi_{i:c\&NE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&NE,0}, \sigma^2_{c\&NE,0})\right] \\
&+ \sum_{i \in O(0,?,1)} I(G_i = n\&EE) \log\left[\pi_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE})\right] \\
&+ \sum_{i \in O(0,?,0)} I(G_i = c\&EN) \log\left[\pi_{i:c\&EN}\right] \\
&+ \sum_{i \in O(0,?,0)} I(G_i = c\&NN) \log\left[\pi_{i:c\&NN}\right] \\
&+ \sum_{i \in O(0,?,0)} I(G_i = n\&NN) \log\left[\pi_{i:n\&NN}\right] \\
&+ \sum_{i \in O(1,1,?)} I(G_i = c\&EE) \log\left[\pi_{i:c\&EE}\right] + \sum_{i \in O(1,1,?)} I(G_i = c\&EN) \log\left[\pi_{i:c\&EN}\right] \\
&+ \sum_{i \in O(1,1,?)} I(G_i = c\&NE) \log\left[\pi_{i:c\&NE}\right] + \sum_{i \in O(1,1,?)} I(G_i = c\&NN) \log\left[\pi_{i:c\&NN}\right] \\
&+ \sum_{i \in O(1,0,?)} I(G_i = n\&EE) \log\left[\pi_{i:n\&EE}\right] + \sum_{i \in O(1,0,?)} I(G_i = n\&NN) \log\left[\pi_{i:n\&NN}\right]
\end{aligned}
$$

where $I(\cdot)$ is the general indicator function. The E-step of the EM algorithm computes the conditional probabilities of each stratum, given the current estimate $\boldsymbol{\theta}^{(t)}_{sci}$, $t = 0, 1, ...$:

• for $i \in O(1, 1, 1)$

$$
P^{(t)}(G_i = c\&EE) = \frac{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1})}{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1}) + \pi^{(t)}_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1})}
$$

$$
P^{(t)}(G_i = c\&EN) = \frac{\pi^{(t)}_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1})}{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1}) + \pi^{(t)}_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1})}
$$

$$
P^{(t)}(G_i = c\&NE) = P^{(t)}(G_i = c\&NN) = P^{(t)}(G_i = n\&EE) = P^{(t)}(G_i = n\&NN) = 0
$$

• for $i \in O(1, 1, 0)$

$$
P^{(t)}(G_i = c\&NE) = \frac{\pi^{(t)}_{i:c\&NE}}{\pi^{(t)}_{i:c\&NE} + \pi^{(t)}_{i:c\&NN}}
$$

$$P^{(t)}(G_i = c\&NN) = \frac{\pi^{(t)}_{i:c\&NN}}{\pi^{(t)}_{i:c\&NE} + \pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = n\&EE) = P^{(t)}(G_i = n\&NN) = 0$$

- for $i \in O(1, 0, 1)$

$$P^{(t)}(G_i = n\&EE) = 1$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = c\&NE)$$

$$= P^{(t)}(G_i = c\&NN) = P^{(t)}(G_i = n\&NN) = 0$$

- for $i \in O(1, 0, 0)$

$$P^{(t)}(G_i = n\&NN) = 1$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = c\&NE)$$

$$= P^{(t)}(G_i = c\&NN) = P^{(t)}(G_i = n\&EE) = 0$$

- for $i \in O(1, ?, 1)$

$$P^{(t)}(G_i = c\&EE) =$$

$$= \frac{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1})}{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1}) + \pi^{(t)}_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1}) + \pi^{(t)}_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

$$P^{(t)}(G_i = c\&EN) =$$

$$= \frac{\pi^{(t)}_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1})}{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1}) + \pi^{(t)}_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1}) + \pi^{(t)}_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

$$P^{(t)}(G_i = n\&EE) =$$

$$= \frac{\pi^{(t)}_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1}) + \pi^{(t)}_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1}) + \pi^{(t)}_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

- for $i \in O(1, ?, 0)$

$$P^{(t)}(G_i = c\&NE) = \frac{\pi^{(t)}_{i:c\&NE}}{\pi^{(t)}_{i:c\&NE} + \pi^{(t)}_{i:c\&NN} + \pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = c\&NN) = \frac{\pi^{(t)}_{i:c\&NN}}{\pi^{(t)}_{i:c\&NE} + \pi^{(t)}_{i:c\&NN} + \pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = n\&NN) = \frac{\pi^{(t)}_{i:n\&NN}}{\pi^{(t)}_{i:c\&NE} + \pi^{(t)}_{i:c\&NN} + \pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = n\&EE) = 0$$

- for $i \in O(0, ?, 1)$

$$P^{(t)}(G_i = c\&EE) =$$

$$= \frac{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,0}, \sigma^{2(t)}_{c\&EE,0})}{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,0}, \sigma^{2(t)}_{c\&EE,0}) + \pi^{(t)}_{i:c\&NE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&NE,0}, \sigma^{2(t)}_{c\&NE,0}) + \pi^{(t)}_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

$$P^{(t)}(G_i = c\&NE) =$$

$$= \frac{\pi^{(t)}_{i:c\&NE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&NE,0}, \sigma^{2(t)}_{c\&NE,0})}{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,0}, \sigma^{2(t)}_{c\&EE,0}) + \pi^{(t)}_{i:c\&NE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&NE,0}, \sigma^{2(t)}_{c\&NE,0}) + \pi^{(t)}_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

$$P^{(t)}(G_i = n\&EE) =$$

$$= \frac{\pi^{(t)}_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}{\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,0}, \sigma^{2(t)}_{c\&EE,0}) + \pi^{(t)}_{i:c\&NE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&NE,0}, \sigma^{2(t)}_{c\&NE,0}) + \pi^{(t)}_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

$$P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = c\&NN) = P^{(t)}(G_i = n\&NN) = 0$$

- for $i \in O(0, ?, 0)$

$$P^{(t)}(G_i = c\&EN) = \frac{\pi^{(t)}_{i:c\&EN}}{\pi^{(t)}_{i:c\&EN} + \pi^{(t)}_{i:c\&NN} + \pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = c\&NN) = \frac{\pi^{(t)}_{i:c\&NN}}{\pi^{(t)}_{i:c\&EN} + \pi^{(t)}_{i:c\&NN} + \pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = n\&NN) = \frac{\pi^{(t)}_{i:n\&NN}}{\pi^{(t)}_{i:c\&EN} + \pi^{(t)}_{i:c\&NN} + \pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&NE) = P^{(t)}(G_i = n\&EE) = 0$$

- for $i \in O(1, 1, ?)$

$$P^{(t)}(G_i = c\&EE) = \frac{\pi^{(t)}_{i:c\&EE}}{\pi^{(t)}_{i:c\&EE} + \pi^{(t)}_{i:c\&EN} + \pi^{(t)}_{i:c\&NE} + \pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = c\&EN) = \frac{\pi^{(t)}_{i:c\&EN}}{\pi^{(t)}_{i:c\&EE} + \pi^{(t)}_{i:c\&EN} + \pi^{(t)}_{i:c\&NE} + \pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = c\&NE) = \frac{\pi^{(t)}_{i:c\&NE}}{\pi^{(t)}_{i:c\&EE} + \pi^{(t)}_{i:c\&EN} + \pi^{(t)}_{i:c\&NE} + \pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = c\&NN) = \frac{\pi^{(t)}_{i:c\&NN}}{\pi^{(t)}_{i:c\&EE} + \pi^{(t)}_{i:c\&EN} + \pi^{(t)}_{i:c\&NE} + \pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = n\&EE) = P^{(t)}(G_i = n\&NN) = 0$$

• for $i \in O(1, 0, ?)$

$$P^{(t)}(G_i = n\&EE) = \frac{\pi^{(t)}_{i:n\&EE}}{\pi^{(t)}_{i:n\&EE} + \pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = n\&NN) = \frac{\pi^{(t)}_{i:n\&NN}}{\pi^{(t)}_{i:n\&EE} + \pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = c\&NE) = P^{(t)}(G_i = c\&NN) = 0$$

The expected log-likelihood $l_E(\boldsymbol{\theta}_{sci}|\mathbf{D}(1), \mathbf{M}_{obs}, \mathbf{S}_{obs}, \mathbf{W}_{obs}, \mathbf{Z}, \mathbf{X})$ is obtained by replacing the $I(G_i = g)$ with the $P^{(t)}(G_i = g)$. The M-step maximizes $l_E(\cdot)$ with respect to $\boldsymbol{\theta}_{sci}$, leading to a new estimate $\boldsymbol{\theta}^{(t+1)}_{sci}$. Iterating this process monotonically increases the likelihood function (6); the algorithm continues until a stopping criterion has been satisfied.

The expected log-likelihood can be decomposed into two parts, one containing the parameters of the wage distribution ($\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$), and the other containing the parameters of the strata membership probabilities ($\boldsymbol{\alpha}$), because

$$\log\left[\pi_{i:g} N_i(\mathbf{X}_i\boldsymbol{\beta}_{g,z}, \sigma^2_{g,z})\right] = \log\left[\pi_{i:g}\right] + \log\left[N_i(\mathbf{X}_i\boldsymbol{\beta}_{g,z}, \sigma^2_{g,z})\right].$$

As a consequence, the two sets of parameters can be updated separately. Standard routines for linear regression (for $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$) and multinomial logistic models (for $\boldsymbol{\alpha}$) can be exploited, weighting the observations with the current probabilities (as estimated in the E-step). For example, updating $\beta_{g,z}$ and $\sigma_{g,z}$ requires maximizing the following function:

$$l_{E:\beta_{g,z},\sigma_{g,z}}(\beta_{g,z}, \sigma_{g,z}|\cdot) = \sum_{i:Z_i=z, S_{i,obs}=1} P^{(t)}(G_i = g) \log\left[N_i(\mathbf{X}_i\boldsymbol{\beta}_{g,z}, \sigma^2_{g,z})\right],$$

with $z = \{0, 1\}$ and $g \in \mathcal{G}$, which is the log-likelihood of a normal model, where each observation is weighted with the current probability of belonging to the principal stratum g. Weighted OLS can then be used for finding $\beta^{(t)}_{g,z}$ and $\sigma^{(t)}_{g,z}$.

EM steps under nonignorable models

The EM algorithm can be easily extended to include nonignorable missing data processes, although this includes the estimation of the parameters governing the missingness mechanism, $\boldsymbol{\theta}_{mis}$.

Under Latent Ignorability, the complete-data log-likelihood function for $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{sci}, \boldsymbol{\theta}_{mis}\}$, given the principal strata, can be written as follows:

$$
\begin{aligned}
&l(\boldsymbol{\theta}|\mathbf{D}(1), \mathbf{M}_{obs}, \mathbf{S}_{obs}, \mathbf{W}_{obs}, \mathbf{G}, \mathbf{Z}, \mathbf{X}) \propto \\
&\sum_{i\in O(1,1,1)} I(G_i = c\&EE) \log\left[\bar{\rho}_{i:c\&EE,1}\pi_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&EE}, \sigma^2_{c\&EE,1})\right] \\
&+ \sum_{i\in O(1,1,1)} I(G_i = c\&EN) \log\left[\bar{\rho}_{i:c\&EN,1}\pi_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&EN,1}, \sigma^2_{c\&EN,1})\right] \\
&+ \sum_{i\in O(1,1,0)} I(G_i = c\&NE) \log\left[\bar{\rho}_{i:c\&NE,1}\pi_{i:c\&NE}\right] + \sum_{i\in O(1,1,0)} I(G_i = c\&NN) \log\left[\bar{\rho}_{i:c\&NN,1}\pi_{i:c\&NN}\right] \\
&+ \sum_{i\in O(1,0,1)} I(G_i = n\&EE) \log\left[\bar{\rho}_{i:n\&EE,1}\pi_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE})\right] \\
&+ \sum_{i\in O(1,0,0)} I(G_i = n\&NN) \log\left[\bar{\rho}_{i:n\&NN,1}\pi_{i:n\&NN}\right] \\
&+ \sum_{i\in O(1,?,1)} I(G_i = c\&EE) \log\left[\bar{\rho}_{i:c\&EE,1}\pi_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&EE,1}, \sigma^2_{c\&EE,1})\right] \\
&+ \sum_{i\in O(1,?,1)} I(G_i = c\&EN) \log\left[\bar{\rho}_{i:c\&EN,1}\pi_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&EN,1}, \sigma^2_{c\&EN,1})\right] \\
&+ \sum_{i\in O(1,?,1)} I(G_i = n\&EE) \log\left[\bar{\rho}_{i:n\&EE,1}\pi_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE})\right] \\
&+ \sum_{i\in O(1,?,0)} I(G_i = c\&NE) \log\left[\bar{\rho}_{i:c\&NE,1}\pi_{i:c\&NE}\right] \\
&+ \sum_{i\in O(1,?,0)} I(G_i = c\&NN) \log\left[\bar{\rho}_{i:c\&NN,1}\pi_{i:c\&NN}\right] \\
&+ \sum_{i\in O(1,?,0)} I(G_i = n\&NN) \log\left[\bar{\rho}_{i:n\&NN,1}\pi_{i:n\&NN}\right] \\
&+ \sum_{i\in O(0,?,1)} I(G_i = c\&EE) \log\left[\bar{\rho}_{i:c\&EE,0}\pi_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&EE,0}, \sigma^2_{c\&EE,0})\right] \\
&+ \sum_{i\in O(0,?,1)} I(G_i = c\&NE) \log\left[\bar{\rho}_{i:c\&NE,0}\pi_{i:c\&NE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{c\&NE,0}, \sigma^2_{c\&NE,0})\right] \\
&+ \sum_{i\in O(0,?,1)} I(G_i = n\&EE) \log\left[\bar{\rho}_{i:n\&EE,0}\pi_{i:n\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}_{n\&EE}, \sigma^2_{n\&EE})\right] \\
&+ \sum_{i\in O(0,?,0)} I(G_i = c\&EN) \log\left[\bar{\rho}_{i:c\&EN,0}\pi_{i:c\&EN}\right] \\
&+ \sum_{i\in O(0,?,0)} I(G_i = c\&NN) \log\left[\bar{\rho}_{i:c\&NN,0}\pi_{i:c\&NN}\right] \\
&+ \sum_{i\in O(0,?,0)} I(G_i = n\&NN) \log\left[\bar{\rho}_{i:n\&NN,0}\pi_{i:n\&NN}\right] \\
&+ \sum_{i\in O(1,1,?)} I(G_i = c\&EE) \log\left[\rho_{i:c\&EE,1}\pi_{i:c\&EE}\right] + \sum_{i\in O(1,1,?)} I(G_i = c\&EN) \log\left[\rho_{i:c\&EN,1}\pi_{i:c\&EN}\right]
\end{aligned}
$$

$$+ \sum_{i \in O(1,1,?)} I(G_i = c\&NE) \log\left[\rho_{i:c\&NE,1}\pi_{i:c\&NE}\right] + \sum_{i \in O(1,1,?)} I(G_i = c\&NN) \log\left[\rho_{i:c\&NN,1}\pi_{i:c\&NN}\right]$$

$$+ \sum_{i \in O(1,0,?)} I(G_i = n\&EE) \log\left[\rho_{i:n\&EE,1}\pi_{i:n\&EE}\right] + \sum_{i \in O(1,0,?)} I(G_i = n\&NN) \log\left[\rho_{i:n\&NN,1}\pi_{i:n\&NN}\right]$$

$$+ \sum_{i \in O(0,?,?)} \sum_{g \in \mathcal{G}} I(G_i = g) \log\left[\rho_{i:g,0}\pi_{i:g}\right]$$

where $\bar{\rho}_{i:g,z} = 1 - \rho_{i:g,z}$. The E-step of the EM algorithm computes the conditional probabilities of each stratum, given the current estimates $\boldsymbol{\theta}^{(t)}$, $t = 0, 1, ...$:

• for $i \in O(1, 1, 1)$

$$P^{(t)}(G_i = c\&EE) = \frac{\bar{\rho}_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1})}{\bar{\rho}_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1}) + \bar{\rho}_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1})}$$

$$P^{(t)}(G_i = c\&EN) = \frac{\bar{\rho}_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1})}{\bar{\rho}_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1}) + \bar{\rho}_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN} N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1})}$$

$$P^{(t)}(G_i = c\&NE) = P^{(t)}(G_i = c\&NN) = P^{(t)}(G_i = n\&EE) = P^{(t)}(G_i = n\&NN) = 0$$

• for $i \in O(1, 1, 0)$

$$P^{(t)}(G_i = c\&NE) = \frac{\bar{\rho}_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE}}{\bar{\rho}_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE} + \bar{\rho}_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = c\&NN) = \frac{\bar{\rho}_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN}}{\bar{\rho}_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE} + \bar{\rho}_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = n\&EE) = P^{(t)}(G_i = n\&NN) = 0$$

• for $i \in O(1, 0, 1)$

$$P^{(t)}(G_i = n\&EE) = 1$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = c\&NE)$$

$$= P^{(t)}(G_i = c\&NN) = P^{(t)}(G_i = n\&NN) = 0$$

• for $i \in O(1, 0, 0)$

$$P^{(t)}(G_i = n\&NN) = 1$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = c\&NE)$$

$$= P^{(t)}(G_i = c\&NN) = P^{(t)}(G_i = n\&EE) = 0$$

• for $i \in O(1, ?, 1)$

$$P^{(t)}(G_i = c\&EE) =$$

$$= \frac{\bar{\rho}_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1})}{\bar{\rho}_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1}) + \bar{\rho}_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1}) + \bar{\rho}_{i:n\&EE,1}\pi^{(t)}_{i:n\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

$$P^{(t)}(G_i = c\&EN) =$$

$$= \frac{\bar{\rho}_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1})}{\bar{\rho}_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1}) + \bar{\rho}_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1}) + \bar{\rho}_{i:n\&EE,1}\pi^{(t)}_{i:n\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

$$P^{(t)}(G_i = n\&EE) =$$

$$= \frac{\bar{\rho}_{i:n\&EE,1}\pi^{(t)}_{i:n\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}{\bar{\rho}_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,1}, \sigma^{2(t)}_{c\&EE,1}) + \bar{\rho}_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EN,1}, \sigma^{2(t)}_{c\&EN,1}) + \bar{\rho}_{i:n\&EE,1}\pi^{(t)}_{i:n\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

• for $i \in O(1, ?, 0)$

$$P^{(t)}(G_i = c\&NE) = \frac{\bar{\rho}_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE}}{\bar{\rho}_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE} + \bar{\rho}_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN} + \bar{\rho}_{i:n\&NN,1}\pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = c\&NN) = \frac{\bar{\rho}_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN}}{\bar{\rho}_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE} + \bar{\rho}_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN} + \bar{\rho}_{i:n\&NN,1}\pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = n\&NN) = \frac{\bar{\rho}_{i:n\&NN,1}\pi^{(t)}_{i:n\&NN}}{\bar{\rho}_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE} + \bar{\rho}_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN} + \bar{\rho}_{i:n\&NN,1}\pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = n\&EE) = 0$$

• for $i \in O(0, ?, 1)$

$$P^{(t)}(G_i = c\&EE) =$$

$$= \frac{\bar{\rho}_{i:c\&EE,0}\pi^{(t)}_{i:c\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,0}, \sigma^{2(t)}_{c\&EE,0})}{\bar{\rho}_{i:c\&EE,0}\pi^{(t)}_{i:c\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,0}, \sigma^{2(t)}_{c\&EE,0}) + \bar{\rho}_{i:c\&NE,0}\pi^{(t)}_{i:c\&NE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&NE,0}, \sigma^{2(t)}_{c\&NE,0}) + \bar{\rho}_{i:n\&EE,0}\pi^{(t)}_{i:n\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

$$P^{(t)}(G_i = c\&NE) =$$

$$= \frac{\bar{\rho}_{i:c\&NE,0}\pi^{(t)}_{i:c\&NE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&NE,0}, \sigma^{2(t)}_{c\&NE,0})}{\bar{\rho}_{i:c\&EE,0}\pi^{(t)}_{i:c\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,0}, \sigma^{2(t)}_{c\&EE,0}) + \bar{\rho}_{i:c\&NE,0}\pi^{(t)}_{i:c\&NE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&NE,0}, \sigma^{2(t)}_{c\&NE,0}) + \bar{\rho}_{i:n\&EE,0}\pi^{(t)}_{i:n\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

$$P^{(t)}(G_i = n\&EE) =$$

$$= \frac{\bar{\rho}_{i:n\&EE,0}\pi^{(t)}_{i:n\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}{\bar{\rho}_{i:c\&EE,0}\pi^{(t)}_{i:c\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&EE,0}, \sigma^{2(t)}_{c\&EE,0}) + \bar{\rho}_{i:c\&NE,0}\pi^{(t)}_{i:c\&NE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{c\&NE,0}, \sigma^{2(t)}_{c\&NE,0}) + \bar{\rho}_{i:n\&EE,0}\pi^{(t)}_{i:n\&EE}N_i(\mathbf{X}_i\boldsymbol{\beta}^{(t)}_{n\&EE}, \sigma^{2(t)}_{n\&EE})}$$

$$P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = c\&NN) = P^{(t)}(G_i = n\&NN) = 0$$

• for $i \in O(0, ?, 0)$

$$P^{(t)}(G_i = c\&EN) = \frac{\bar{\rho}_{i:c\&EN,0}\pi^{(t)}_{i:c\&EN}}{\bar{\rho}_{i:c\&EN,0}\pi^{(t)}_{i:c\&EN} + \bar{\rho}_{i:c\&NN,0}\pi^{(t)}_{i:c\&NN} + \bar{\rho}_{i:n\&NN,0}\pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = c\&NN) = \frac{\bar{\rho}_{i:c\&NN,0}\pi^{(t)}_{i:c\&NN}}{\bar{\rho}_{i:c\&EN,0}\pi^{(t)}_{i:c\&EN} + \bar{\rho}_{i:c\&NN,0}\pi^{(t)}_{i:c\&NN} + \bar{\rho}_{i:n\&NN,0}\pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = n\&NN) = \frac{\bar{\rho}_{i:n\&NN,0}\pi^{(t)}_{i:n\&NN}}{\bar{\rho}_{i:c\&EN,0}\pi^{(t)}_{i:c\&EN} + \bar{\rho}_{i:c\&NN,0}\pi^{(t)}_{i:c\&NN} + \bar{\rho}_{i:n\&NN,0}\pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&NE) = P^{(t)}(G_i = n\&EE) = 0$$

• for $i \in O(1, 1, ?)$

$$P^{(t)}(G_i = c\&EE) = \frac{\rho_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE}}{\rho_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE} + \rho_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN} + \rho_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE} + \rho_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = c\&EN) = \frac{\rho_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN}}{\rho_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE} + \rho_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN} + \rho_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE} + \rho_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = c\&NE) = \frac{\rho_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE}}{\rho_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE} + \rho_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN} + \rho_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE} + \rho_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = c\&NN) = \frac{\rho_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN}}{\rho_{i:c\&EE,1}\pi^{(t)}_{i:c\&EE} + \rho_{i:c\&EN,1}\pi^{(t)}_{i:c\&EN} + \rho_{i:c\&NE,1}\pi^{(t)}_{i:c\&NE} + \rho_{i:c\&NN,1}\pi^{(t)}_{i:c\&NN}}$$

$$P^{(t)}(G_i = n\&EE) = P^{(t)}(G_i = n\&NN) = 0$$

• for $i \in O(1, 0, ?)$

$$P^{(t)}(G_i = n\&EE) = \frac{\rho_{i:n\&EE,1}\pi^{(t)}_{i:n\&EE}}{\rho_{i:n\&EE,1}\pi^{(t)}_{i:n\&EE} + \rho_{i:n\&NN,1}\pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = n\&NN) = \frac{\rho_{i:n\&NN,1}\pi^{(t)}_{i:n\&NN}}{\rho_{i:n\&EE,1}\pi^{(t)}_{i:n\&EE} + \rho_{i:n\&NN,1}\pi^{(t)}_{i:n\&NN}}$$

$$P^{(t)}(G_i = c\&EE) = P^{(t)}(G_i = c\&EN) = P^{(t)}(G_i = c\&NE) = P^{(t)}(G_i = c\&NN) = 0$$

• for $i \in O'(0,0)$

$$P^{(t)}(G_i = g) = \frac{\rho^{(t)}_{i:g,0}\pi^{(t)}_{i:g}}{\sum_{g\in\mathcal{G}}\rho^{(t)}_{i:g,0}\pi^{(t)}_{i:g}}$$

The expected log-likelihood, $l_E(\boldsymbol{\theta}|\mathbf{D}(1), \mathbf{M}_{obs}, \mathbf{S}_{obs}, \mathbf{W}_{obs}, \mathbf{Z}, \mathbf{X})$, is obtained by replacing $I(G_i = g)$ with $P^{(t)}(G_i = g)$. The M-step maximizes $l_E(\cdot)$ with respect to $\boldsymbol{\theta}$, to obtain $\boldsymbol{\theta}^{(t+1)}$.

	Treatment			Control			Difference	
Variable	Prop. non-miss.	Mean	Std. Dev.	Prop. non-miss.	Mean	Std. Dev.	Diff.	Std. Err.
Female	1.00	0.41	0.49	1.00	0.41	0.49	0.00	0.01
Age at baseline (yrs)	1.00	18.85	2.17	1.00	18.79	2.13	0.05	0.04
White	1.00	0.27	0.44	1.00	0.26	0.44	0.01	0.01
With a partner	0.98	0.06	0.24	0.97	0.06	0.24	0.00	0.00
Has children	0.99	0.18	0.38	0.99	0.18	0.38	0.00	0.01
Education (yrs of schooling)	0.98	10.07	1.53	0.97	10.08	1.51	−0.01	0.03
Ever arrested	0.98	0.26	0.44	0.98	0.26	0.44	0.00	0.01
Mother's education (yrs of schooling)	0.80	11.52	2.56	0.78	11.54	2.61	−0.02	0.05
Father's education (yrs of schooling)	0.60	11.46	2.87	0.59	11.55	2.86	−0.09	0.06
Household Inc. > 6000	0.62	0.55	0.50	0.63	0.54	0.50	0.01	0.01
Personal Inc. > 6000	0.91	0.09	0.28	0.91	0.08	0.27	0.01	0.01
At baseline:								
Have Job	0.96	0.21	0.41	0.96	0.21	0.41	0.00	0.01
Had Job, prev. yr.	0.98	0.65	0.48	0.98	0.64	0.48	0.01	0.01
Months empl., prev. yr.	0.93	3.79	4.27	0.93	3.77	4.30	0.02	0.08
Earnings, prev. yr. (US dollars)	0.91	2904.89	4529.84	0.91	2867.17	4420.10	37.72	82.06
N	8688			5299				

Table 1: Univariate descriptive statistics for pre-treatment covariates by treatment group, computed using units with observed values for the specified variable and using design weights.

	Treatment			Control			Difference	
Variable	Prop. non-miss.	Mean	Std. Dev.	Prop. non-miss.	Mean	Std. Dev.	Diff.	Std. Err.
Enrolled in Job Corps within 6 months from assignment	0.99	0.68	0.47	–	–	–	–	–
Week 52								
Employed	0.97	0.38	0.48	0.96	0.44	0.50	−0.06	0.01
Weekly earnings	0.97	98.78	164.86	0.96	109.03	162.63	−10.25	2.91
Weekly hours	0.97	15.86	22.69	0.96	18.24	23.03	−2.38	0.40
Wage	0.37	6.20	3.10	0.43	5.93	2.72	0.28	0.02
Week 130								
Employed	0.98	0.51	0.50	0.98	0.49	0.50	0.02	0.01
Weekly earnings	0.98	167.72	221.40	0.98	153.14	202.65	14.58	3.77
Weekly hours	0.98	22.74	24.91	0.98	21.60	24.62	1.14	0.44
Wage	0.50	7.37	3.53	0.48	7.03	2.94	0.34	0.02
Week 208								
Employed	0.77	0.61	0.49	0.82	0.57	0.50	0.04	0.01
Weekly earnings	0.77	228.64	254.43	0.82	202.82	232.66	25.82	4.79
Weekly hours	0.77	27.38	25.03	0.82	25.24	24.95	2.14	0.49
Wage	0.47	8.30	3.94	0.47	8.06	3.77	0.24	0.03

Table 2: Univariate descriptive statistics for outcome variables by treatment group, computed using units with observed values for the specified variable and using design weights.

Observed subgroups $O(Z_i, D_i(1), S_{i,obs})$	Latent strata
$O(1,1,1) = \{i : Z_i = 1, D_i(1) = 1, S_{i,obs} = 1\}$	$c\&EE$, $c\&EN$
$O(1,1,0) = \{i : Z_i = 1, D_i(1) = 1, S_{i,obs} = 0\}$	$c\&NN$, $c\&NE$
$O(1,0,1) = \{i : Z_i = 1, D_i(1) = 0, S_{i,obs} = 1\}$	$n\&EE$
$O(1,0,0) = \{i : Z_i = 1, D_i(1) = 0, S_{i,obs} = 0\}$	$n\&NN$
$O(1,?,1) = \{i : Z_i = 1, D_i(1) = ?, S_{i,obs} = 1\}$	$n\&EE$, $c\&EE$, $c\&EN$
$O(1,?,0) = \{i : Z_i = 1, D_i(1) = ?, S_{i,obs} = 0\}$	$n\&NN$, $c\&NN$, $c\&NE$
$O(0,?,1) = \{i : Z_i = 0, D_i(1) = ?, S_{i,obs} = 1\}$	$c\&EE$, $c\&NE$, $n\&EE$
$O(0,?,0) = \{i : Z_i = 0, D_i(1) = ?, S_{i,obs} = 0\}$	$c\&EN$, $c\&NN$, $n\&NN$
$O(1,1,?) = \{i : Z_i = 1, D_i(1) = 1, S_{i,obs} = ?\}$	$c\&EE$, $c\&EN$, $c\&NE$, $c\&NN$
$O(1,0,?) = \{i : Z_i = 1, D_i(1) = 0, S_{i,obs} = ?\}$	$n\&EE$, $n\&NN$
$O(0,?,?) = \{i : Z_i = 0, D_i(1) = ?, S_{i,obs} = ?\}$	$c\&EE$, $c\&EN$, $c\&NE$, $c\&NN$, $n\&EE$, $n\&NN$

Table 3: Correspondence between observed subgroups and latent strata

week	$\pi_{c\&EE}$	$\pi_{c\&EN}$	$\pi_{c\&NE}$	$\pi_{c\&NN}$	$\pi_{n\&EE}$	$\pi_{n\&NN}$	$\Delta^{(ZS)}$	$\Delta^{(DS)}$	$\Delta^{(DW)}$	λ_M	λ_{0W}	λ_{S0}
52	0.236	0.032	0.049	0.397	0.127	0.159	−0.017	−0.024	0.276	8.61	1.03	3.67
130	0.293	0.067	0.052	0.298	0.139	0.151	0.015	0.022	0.247	8.06	1.36	2.44
208	0.377	0.044	0.035	0.261	0.162	0.120	0.009	0.013	0.290	4.89	0.92	2.26

Table 4: Maximum likelihood estimates of the average effects of treatment assignment on employment ($\Delta^{(ZS)}$) and of the average treatment effects on employment for compliers ($\Delta^{(DS)}$) and on wages for always-employed compliers ($\Delta^{(DW)}$), at weeks 52, 130 and 208. For each week, we provide the estimated proportions in the principal strata; λ_M is the scaled LRT statistic if the null model assumes monotonicity of employment; λ_{S0}, $\lambda_{(0W)}$ represent the scaled LRT statistics for the null model with constraints $\Delta^{(DW)} = 0$ and $\Delta^{(DS)} = 0$, respectively.

week	$\bar{W}_{c\&EE,0}$	(s.e.)	$\bar{W}_{c\&EE,1}$	(s.e.)	$\bar{W}_{c\&EN,1}$	(s.e.)	$\bar{W}_{c\&NE,0}$	(s.e.)	$\bar{W}_{n\&EE}$	(s.e.)
52	5.52	(0.000)	5.80	(0.000)	7.32	(0.015)	6.80	(0.030)	6.51	(0.001)
130	6.44	(0.000)	6.69	(0.000)	9.22	(0.009)	7.22	(0.022)	7.94	(0.001)
208	7.47	(0.001)	7.76	(0.001)	9.27	(0.026)	8.99	(0.096)	8.97	(0.001)

Table 5: Maximum likelihood estimates of the average wages in USD at weeks 52, 130 and 208 (asymptotic standard errors in parentheses).

	Variable	*c&EE*	*c&EN*	*c&NE*	*c&NN*	*n&EE*	*n&NN*
Week 52	Female	0.41	0.26	0.25	0.44	0.40	0.45
	Age at baseline	18.9	19.0	19.3	18.4	19.5	18.8
	White	0.34	0.40	0.34	0.20	0.33	0.23
	With a partner	0.07	0.03	0.04	0.04	0.10	0.08
	Has children	0.17	0.09	0.11	0.17	0.21	0.25
	Education	10.2	10.2	10.1	9.8	10.6	9.9
	Ever arrested	0.24	0.29	0.32	0.24	0.28	0.31
	Mother's education	11.73	11.68	11.64	11.41	11.63	11.51
	Fathers' education	11.68	12.11	11.69	11.41	11.60	11.51
	Household income > 6000	0.58	0.61	0.58	0.47	0.60	0.48
	Personal income > 6000	0.10	0.14	0.07	0.05	0.14	0.06
	Have job	0.29	0.29	0.30	0.14	0.32	0.16
	Had Job, prev. yr.	0.75	0.76	0.72	0.55	0.80	0.58
	Months in Job, prev. yr.	4.97	5.08	5.06	2.83	5.57	3.08
	Earnings, prev. yr.	3889.6	4112.4	4379.8	1973.4	4780.8	2508.2

Table 6: Estimated means of covariates within principal strata, computed using design weights and estimated membership probabilities, week 52.

	Variable	*c&EE*	*c&EN*	*c&NE*	*c&NN*	*n&EE*	*n&NN*
Week 130	Female	0.42	0.25	0.19	0.47	0.40	0.46
	Age at baseline	18.95	18.88	18.99	18.36	19.37	18.89
	White	0.30	0.36	0.38	0.19	0.31	0.24
	With a partner	0.05	0.06	0.06	0.04	0.08	0.10
	Has children	0.16	0.14	0.15	0.16	0.22	0.25
	Education	10.20	10.10	10.05	9.81	10.46	9.96
	Ever arrested	0.23	0.29	0.32	0.25	0.28	0.31
	Mother's education	11.54	11.53	11.69	11.51	11.55	11.59
	Fathers' education	11.51	11.82	11.99	11.46	11.53	11.55
	Household income > 6000	0.54	0.62	0.60	0.46	0.59	0.49
	Personal income > 6000	0.08	0.11	0.10	0.05	0.12	0.08
	Have job	0.25	0.27	0.24	0.15	0.28	0.19
	Had Job, prev. yr.	0.69	0.74	0.70	0.55	0.76	0.60
	Months in Job, prev. yr.	4.35	5.03	4.20	2.88	5.04	3.39
	Earnings, prev. yr.	3221.18	4112.41	3755.92	2062.57	4290.66	2775.56

Table 7: Estimated means of covariates within principal strata, computed using design weights and estimated membership probabilities, week 130.

	Variable	*c&EE*	*c&EN*	*c&NE*	*c&NN*	*n&EE*	*n&NN*
Week 208	Female	0.39	0.28	0.29	0.47	0.41	0.45
	Age at baseline	18.93	18.42	18.54	18.45	19.31	18.85
	White	0.29	0.44	0.41	0.18	0.30	0.23
	With a partner	0.05	0.04	0.08	0.04	0.10	0.08
	Has children	0.17	0.10	0.19	0.17	0.22	0.23
	Education	10.17	10.01	9.93	9.79	10.41	9.94
	Ever arrested	0.23	0.33	0.35	0.26	0.27	0.32
	Mother's education	11.57	11.72	11.59	11.44	11.61	11.55
	Fathers' education	11.54	11.96	11.44	11.50	11.60	11.50
	Household income > 6000	0.54	0.71	0.73	0.43	0.57	0.49
	Personal income > 6000	0.09	0.09	0.06	0.04	0.12	0.07
	Have job	0.25	0.28	0.19	0.14	0.27	0.17
	Had Job, prev. yr.	0.71	0.71	0.72	0.51	0.72	0.62
	Months in Job, prev. yr.	4.47	4.17	4.21	2.67	4.84	3.34
	Earnings, prev. yr.	3354.86	3666.80	3622.23	1884.33	4067.85	2731.00

Table 8: Estimated means of covariates within principal strata, computed using design weights and estimated membership probabilities, week 208.

	MAR					LI and (4)					LI and (5)				
week	$\Delta^{(DS)}$	$\Delta^{(DW)}$	λ_M	λ_{S0}	λ_{0W}	$\Delta^{(DS)}$	$\Delta^{(DW)}$	λ_M	λ_{S0}	λ_{0W}	$\Delta^{(DS)}$	$\Delta^{(DW)}$	λ_M	λ_{S0}	λ_{0W}
52	−0.024	0.276	8.61	3.67	1.03	−0.017	0.268	10.12	3.79	1.86	−0.016	0.263	8.65	3.69	1.93
130	0.022	0.247	8.06	2.44	1.36	0.022	0.252	7.99	2.43	1.33	0.023	0.246	8.09	2.35	1.36
208	0.013	0.290	4.89	2.26	0.92	0.009	0.303	4.72	2.39	0.87	0.008	0.278	4.97	2.15	0.87

Table 9: Maximum likelihood estimates of average treatment effects and scaled LRT statistics under different assumptions about the missingness mechanism.