A Nonparametric Finite Mixture Approach to Difference-in-Difference Estimation, with an Application to On-the-job Training and Wages

Oliver Cassagneau-Francis, Robert Gary-Bobo, Julie Pernaudet, Jean-Marc Robin

A Nonparametric Finite Mixture Approach to Difference-in-Difference Estimation, with an Application to On-the-job Training and Wages

Oliver Cassagneau-Francis* Robert Gary-Bobo† Julie Pernaudet‡
Jean-Marc Robin§

October 10, 2022

Abstract

We develop a finite-mixture framework for nonparametric difference-in-difference analysis with unobserved heterogeneity correlating treatment and outcome. Our framework includes an instrumental variable for the treatment, and we demonstrate that this allows us to relax the common-trend assumption. Outcomes can be modeled as first-order Markovian, provided at least 2 post-treatment observations of the outcome are available. We provide a nonparametric identification proof. We apply our framework to evaluate the effect of on-the-job training on wages, using novel French linked employee-employer data. Estimating our model using an EM-algorithm, we find small ATEs and ATTs on hourly wages, around 1%.

Keywords: Finite Mixtures; Unobserved Heterogeneity; EM Algorithm, Wage Distributions; Training; Matched Employer-Employee Data.
JEL codes: E24; E32; J63; J64

*University College London; E-mail: o.cassagneau-francis@ucl.ac.uk
†Université Paris 1 Panthéon-Sorbonne, CES, and CREST. E-mail: garybobo@univ-paris1.fr
‡University of Chicago; E-mail: jpernaudet@uchicago.edu
§Sciences Po, Paris; E-mail jeanmarc.robin@sciencespo.fr; Sciences-Po, Department of Economics, 28 rue des St Pères, 75007 Paris, France. I gratefully acknowledge the support from the European Research Council (grant reference ERC-2020-ADG-101018130).

1 Introduction

Differences-in-differences (DiD) is a widely used method to estimate treatment effects in applied economics. The conventional approach compares the average outcome of a treatment group to the average outcome of a control group before and after the program of interest is implemented. For the DiD estimator to identify the causal effect of the program, researchers need to assume that in the absence of treatment, the outcome of both groups would have followed parallel trends over time (*common trend assumption).* However, in practice, the common trend assumption is often violated. Recent papers have relaxed this assumption to allow it to hold only conditional on observable pre-treatment covariates (see Sant'Anna and Zhao, 2020, for an analysis of the properties of DiD estimators in this case). While conditioning on observables makes the common trend assumption less restrictive, researchers do not necessarily observe all the variables needed to capture all possible confounders and apply this framework plausibly.

In this paper, we suggest a novel extension to the aforementioned approach by allowing for *unobserved* heterogeneity in a DiD framework. Borrowing from a method traditionally used in structural modeling, we relax the canonical common trend assumption to allow it to hold conditional on latent types that capture any possible sources of heterogeneity. We demonstrate that nonparametric identification is achieved by using an excluded variable that affects the treatment probability without directly affecting outcomes, similar to an instrumental variable. However, while finding an instrument that satisfies the standard exclusion restriction is typically challenging, in our framework, this exclusion restriction needs to hold only conditional on the latent types. Further, we show that the conventional monotonicity assumption is not needed in our model. Our framework thus allows for a broad set of candidates to be considered for the excluded variable and offers wide applicability.

Leveraging linked employee-employer survey data matched with administrative data on wages, we apply our model to the estimation of the impact of job training on wages. Our instrument is whether or not the worker has received information about training opportunities. It is a variable that affects the treatment probability for each type, potentially in a different way. As we observe wages more than twice, we allow for the outcome process to be autoregressive by modeling it as a Markovian process.

The availability of repeated observations of the outcome variable is crucial for identification, all the more so depending on the assumed dynamics. The benefits of panel data in difference-in-difference contexts are studied in Bonhomme and Sauder (2011); Freyaldenhoven et al. (2019) and in Callaway and Li (2019); Li and Li (2019); Sant'Anna and Zhao (2020). These papers maintain a common trend assumption, except for the first one. As far as we know, Bonhomme and Sauder (2011) is the only paper that replaces a standard common trend assumption by a structural assumption on the way unobserved heterogene-

ity determines outcomes. Specifically, they assume a linear factor structure and solve the (semiparametric) identification problem using nonparametric deconvolution techniques.[1] Freyaldenhoven et al. (2019) share the factor structure of Bonhomme and Sauder's framework and some identification ideas.[2] We depart from the linear factor structure, allowing for instance unobserved heterogeneity to condition outcome variances. Some restriction on the distribution of latent types is however necessary. We assume that there exist a finite number of groups and that outcomes and treatments are drawn from a distribution that is specific to each group. By giving up continuity,[3] we gain more flexibility and also a simpler method of identification based on standard matrix algebra.

Under the assumptions about discrete heterogeneity and the instrument described above, we prove identification of Average Treatment Effects (ATEs) that are conditional on the unobserved types, as well as heterogeneous treatment probabilities. We also show that each standard difference-in-difference estimator obtained by applying OLS or IV estimation procedures to the wage panel equations is the sum of different weighted means of conditional average treatment effects (ATT and LATE), plus a bias reflecting the violation of the common trend assumption.

After proving identification, we estimate a flexible parametric specification using the (sequential) Expectation-Maximization (EM) algorithm. Standard errors are obtained by bootstrap. The results show that treatment effects vary with type, but once aggregated they are very small and insignificant. All three ways of aggregating conditional ATEs (aggregate ATE, ATT, and LATE) yield similar estimates of around 1%. We conclude that on-the-job training has no or a very limited effect on wages. The biases resulting from heterogeneous trends are found to be of a similar order of magnitude to the aggregate treatment effects. We also find a sizable share of the bias on the IV estimator of around 1%-2% that reflects small-sample deviations from assumed restrictions in the population.

The rest of this paper proceeds as follows. We first end the introduction with a discussion of the literature on training. Then, section 2 presents the model, the associated nonparametric identification result as well as the links between our model's estimates and ATT, ATE, and other parameters of interest. Section 3 describes our dataset and presents a preliminary econometric analysis using standard econometric methods. Section 4 presents and discusses the estimation results. We conclude in section 5.

[1] More precisely, the special case studied in Section II.B does satisfy the common trend assumption, since, by taking differences in outcomes, the fixed effect disappears. In Section II.C, they allow for different factor loadings on the latent factor, but these factor loadings are assumed independent of the treatment, which comes close to a common trend assumption.

[2] The pre-treatment periods of Freyaldenhoven *et al.* play a similar role as the "instrument" of Bonhomme and Sauder, as far as the identification of factor loadings is concerned.

[3] As for example in Hu and Schennach (2008); Allman et al. (2009); Kasahara and Shimotsu (2009a); Hu and Shum (2012); Shiu and Hu (2013); Henry et al. (2014); Hu (2015); Sasaki (2015); Bonhomme et al. (2016a,b, 2017a,b, 2019).

Literature on training. The literature on the effect of training (and active labor market programs) is huge. Fialho et al. (2019) provide the most recent survey and exhaustive evaluation of the different forms of adult learning — informal (on the job), non-formal and formal (depending on whether the institution providing training is public or not) — for various countries. The effect of non-formal training on wages is estimated between 13% and 30%, with and without controls. When a control-function estimator is used, the estimated effect of training remains high, around 11% on average, but with a wide range across countries. Before Fialho et al. (2019), several other authors had reviewed this literature (see Heckman et al. (1999); McCall et al. (2016) and the meta-analyses of Card et al. (2010, 2018) and Haelermans and Borghans (2012)). See also the classic paper by LaLonde (1986). The estimated impacts of training on wages and productivity are generally found to be positive; the effects on the risk of unemployment are often ambiguous.

Many of the contributions devoted to training programs are based on non-experimental data with a panel structure and rely on fixed-effects estimators. Fixed-effects approaches are used in the pioneering work of Ashenfelter (1978), in the contributions of (among many others) Lynch (1992), on NLSY data; Booth (1993); Blundell et al. (1999), both on British data; Krueger and Rouse (1998), on American firm-level data; Pischke (2001), on German GSOEP data; Schoene (2004), on Norwegian data.

Few papers rely on instrumental variables, maybe because it is difficult to find convincing instruments for participation in training programs (yet, see Bartel (1995); Parent (1999); Abadie et al. (2002)). Some contributions controlled for selection in training using Heckman's two-stage estimator (*e.g.* LaLonde (1986); Booth (1993); Goux and Maurin (2000)). A behavioral approach to training participation is explored in Caliendo et al. (2016). Other contributions use matching estimators (Brodaty et al., 2001; Gerfin and Lechner, 2002; Kluve et al., 2012).

A number of recent papers follow Abadie et al. (2002) and use randomized trials; see *e.g.* Lee (2009), Attanasio et al. (2011); Grip and Sauermann (2012); Ba et al. (2017), Sandvik et al. (2021). The importance of the comparison group construction is illustrated by Leuven and Oosterbeek (2008). They narrow down their comparison group to pick only "workers who [were] willing to undertake training and whose employers [were] prepared to provide it, but did not attend the training course they wanted, due to some random event" (Leuven and Oosterbeek, 2008, p. 426). This strict choice of comparison group reduces the estimated coefficient on training to almost zero, down from between 5–15% for less restrictive choices.[4]

Most papers consider the impact of training on wages *and* productivity. Human capital theory suggests that, under conditions of perfect competition, employers should refuse to pay for training. At least, they would refuse to finance general training, which is typically

[4]On this point, see also Sandvik et al. (2021).

portable, and would allow workers to quit the firm and find a job with a higher wage. But under imperfectly competitive conditions, in particular, under asymmetric information about workers' abilities, it can be shown that the firm should be willing, either to subsidize training, or to share the benefits of training with the worker, (see Acemoglu and Pischke, 1998, 1999). A number of papers use wage equations and production functions to test this prediction and do indeed find positive effects on both productivity and wages.[5]

There also exists a literature on transition and duration models, studying the effects of training on the duration of employment and unemployment spells (see Ridder (1986), on Dutch data; Gritz (1993), on NLSY data; Bonnal et al. (1997), on French data; Crepon et al. (2009), using methods developed in Abbring and Berg (2003)).

Finally, an important question is to assess the importance and effects of unobserved heterogeneity, as well as the dynamic structure of the treatment effects of training (for recent progress on these two fronts, see Rodriguez et al. (2018)). Our paper addresses these questions within a nonparametric DiD framework that we describe below.

2 The model

We frame the model in terms of the application we are interested in (the effect of training on wages) but the methodology could be used in many other setups.

We study a population of N workers indexed by i. The outcome variable is the worker's nominal hourly wage (in logs). It is denoted w_{it} and is observed at the end of three consecutive years indexed by $t = 1, 2, 3$. Some workers engage in a training session after the first wage observation, in which case $d_i = 1$, and $d_i = 0$ otherwise. Wage w_{i1} is observed before training, and w_{i2}, w_{i3} are observed after training (if training takes place at all). Our method works with more wage observations before training and just one observation after training. As we shall later explain, we use three wage observations instead of two in order to identify autoregressive wage dynamics.

Our goal is to measure the causal impact of training on wages in periods $t = 2, 3$. We assume that treatment d_i is a binary variable, although the model and the proof of identification encompass the case of a treatment variable with any finite number of values. Specifically, we could allow for different types of training, for example by duration. Moreover, there are two observations of the outcome variable after treatment because we want to allow for the possibility that the outcome is a Markovian (autoregressive) process.

We single out, from all potential control variables, a variable $z_i \in \{0, 1\}$, indicating if the worker reports receiving information about the availability of training sessions through any of the following channels: hierarchy, human resources, coworkers, or unions. This variable will be used as an instrument for the selection into treatment.

[5] See Ballot et al. (2006), Dearden et al. (2006); Konings and Vanormelingen (2015).

We assume that workers can be clustered into a finite number H of *unobserved* groups: $h \in \{1, ..., H\}$. The distribution of all variables w_{it}, z_i and d_i, including the instrument, potentially varies across latent groups. We think of these latent groups as embodying all the heterogeneity, such as education, health, experience, that is observed and unobserved and which conditions wages and training. However, it is of course possible to first cluster the data, say by education, and run the study separately within each education group.[6]

Semi-parametric versions of our model can also easily be worked out, at the cost of restrictions on the interaction between observed and unobserved characteristics. In the application, we will classify workers from observations (w_{it}, d_i, z_i) and examine the correlations between the estimated classification h and a set of available controls, *ex post*.

We start by making the following assumption on wages.

Assumption 1 (Wage process)**.** *The wage process is first-order Markov and bounded, and is independent of the instrument given type and treatment.*

As already noted, it is not necessary at all that the outcome process be dynamic. If wages are iid, then a single wage observation after treatment is sufficient for identification. However, if the outcome process is dynamic given unobserved heterogeneity, then even two wage observations are not enough (i.e. one before and one after treatment). This is expected; see for example Kasahara and Shimotsu (2009b); Hu and Shum (2012). Three wage observations are the minimum amount of data that we require for identification. More observations will always be better — though this might introduce complications regarding the timing of treatment. These three wage observations can be such that one is observed before treatment and two are observed after, as in our empirical setup; or we could observe two wages before treatment and one after. Lastly, it is easy to extend the framework and allow for an autoregressive process of higher degree; but the Markov property is essential. So, we cannot for the moment allow for a non-Markovian process such as the sum of a random walk and an iid innovation. Without the bounded support assumption, our identification method will only allow us to identify parameters up to any given error.

Assumption 1 is also the first restriction imposed on the special variable z_i. The variable z_i is a valid instrument for training insofar as it does not affect wages once heterogeneity and training are controlled for. This exclusion restriction is fundamental for our identification result. In our empirical application, the instrument is whether the worker had access to information on training. Conditional on all worker heterogeneity, it seems reasonable to assume that training information has no causal effect on wages. Testing this exclusion restriction is difficult because individual types are latent variables, which will be recovered only under this (and other) assumptions.

[6] We tried pre-clustering the data along a number of observed dimensions (age, education, gender, firm size) to little effect. Our final estimation strategy does pre-cluster the data, though only on wages to capture observed and unobserved heterogeneity.

Let $F_t(w_{it}|h,d)$ denote the distribution function for the marginal distribution of wages w_{it} given treatment and type. We use a lower-case f to denote the corresponding probability mass or density functions. Let $F_{t|s}(w_{it}|w_{is},h,d)$ denote the distribution function for the conditional distribution of w_{it} given w_{is} (we use $s = t \pm 1$). Let $\mathcal{W}_2(h,d)$ be the support of $f_2(w_2|h,d)$ (i.e. the set of wages w such that $f_2(w_2|h,d) > 0$) and let $\mathcal{W}_2(d) = \bigcup_h \mathcal{W}_2(h,d)$ be the joint support (different types may have different supports). Lastly, let $\pi(h,z,d)$ denote the probability mass of workers of type $h \in \{1,...,H\}$, with values of the instrument $z \in \{0,1\}$ and of treatment $d \in \{0,1\}$.

Our framework relates to the standard difference-in-differences model, since we compare pre- and post-treatment wages. It can also be interpreted as a version of the Roy model used by Carneiro, Heckman and Vytlacil in a number of papers (see Heckman and Vytlacil, 2005; Carneiro and Lee, 2009; Carneiro et al., 2010, 2011 for example). The main difference is that we explicitly model the dependence between error terms via the latent factor h. Specifically, a possible interpretation of our model combines an outcome and a choice (or selection) equation as follows. Let $y(0)$, $y(1)$ denote the potential outcomes (i.e., post-treatment wages) for $d = 0$ or $d = 1$ and let $c(h,z) + v$ be a random training cost depending on h and z. Then, a standard Roy model would have $d = 1$ if and only if the expected return $\mathrm{E}[y(1) - y(0) \,|\, h]$ is greater than the cost $c(h,z) + v$.

2.1 Identification

In this section, we describe the conditions under which our model is identified.

All the relevant information is in the likelihood of the information available at the individual level, namely the instrument z, the treatment d, and the three wages w_1 (before treatment) and w_2, w_3 (after treatment):

$$p(z,d,w_1,w_2,w_3) = \sum_h \pi(h,z,d)\, f_2(w_2|h,d)\, f_{1|2}(w_1|w_2,h,d)\, f_{3|2}(w_3|w_2,h,d). \qquad (1)$$

Notice how we condition the densities of w_{i1} and w_{i3} on w_{i2}. The static case of iid wages given latent types and treatment can be seen as a particular case where $\mathcal{W}_2(d)$ is reduced to a singleton.

We show that all the components of the right-hand side of equation (1) are identified under the following assumptions.

Assumption 2 (Overlap)**.** *For all h, d, $\pi(h,0,d) \neq 0$.*

Assumption 2 is standard and means that workers of all types have a positive probability of being both treated and non-treated for at least one instrument value, arbitrarily set equal to zero.

Assumption 3 (Linear independence). *For all d, $w_2 \in \mathcal{W}_2(d)$, and $t = 1, 3$, there exists grids of wages (w_t) such that the systems $\left\{F_{t|2}(w_t|w_2, h, d), \forall h : f_2(w_2|h, d) \neq 0\right\}$ are linearly independent.*

Any latent type such that its conditional wage distribution can be replicated as a linear combination of the other types' distributions cannot be separately identified from the other types. This is also a standard assumption.

Assumption 4 (Rank condition). *For all d, and all $h \neq h'$, $\dfrac{\pi(h,1,d)}{\pi(h,0,d)} \neq \dfrac{\pi(h',1,d)}{\pi(h',0,d)}$.*

This assumption is standard in the literature on latent group identification. For example, it is related to Assumption 3 in Bonhomme et al. (2019) and Assumption 2.3 in Hu (2008) (or Assumption 3 in Hu's (2017) survey). Assumption 4 requires different exposures to the instrument for all types, whatever the treatment. If the ratios $\frac{\pi(h,1,d)}{\pi(h,0,d)}$ differ by h, then we can as well relabel types so that $\frac{\pi(h,1,d)}{\pi(h,0,d)}$ is increasing in h. So, given d, we require probabilities $\pi(h, z, d)$ to be log-supermodular. In other words, the instrument must predict individual types given treatment; but it must not predict wages given type and treatment.

In our empirical application, the instrument is training information. Conditional on all worker heterogeneity, it seems reasonable to assume that it has no causal effect on wages (Assumption 1) but receiving information about training opportunities should increase the probability to be treated. The instrument is an intention to treat. Now, with more transparent notation,

$$\frac{\pi(h,1,d)}{\pi(h,0,d)} = \frac{\Pr(d|h, z=1)\Pr(z=1|h)\pi(h)}{\Pr(d|h, z=0)\Pr(z=0|h)\pi(h)}. \tag{2}$$

Assumption 4 can thus hold because different types show different probabilities of complying, or because different types show different intentions to treat.

Finally, we assume that the pre-treatment wage distribution should be independent of the treatment. Hence, pre-treatment wages must be independent of both the instrument (Assumption 1) and the treatment.

Assumption 5 (Predetermination). *For all types h, $f_1(w_1|h, d) = f_1(w_1|h)$ and $f_1(\cdot|h) \neq f_1(\cdot|h')$ for all h, h'.*

This assumption is useful to recover a common labelling of groups across treatments. Predetermination does not always hold (even conditional on all relevant heterogeneity). For example, an "Ashenfelter dip" (wages drop before treatment) could be observed if employers make workers pay for the forthcoming training. In our application, most training sessions are rather short (a few days, rarely a whole week) and the pre-treatment wage is observed a full year before training.

Theorem 1 (Identification given (d, w_2))**.** *Under Assumptions 1-5, the number of latent groups H, and the functional parameters $\pi(h, z, d)$, $f_2(w_2|h, d)$, $F_{1|2}(w_1|w_2, h, d)$, and $F_{3|2}(w_3|w_2, h, d)$ are identified up to labelling. Using Bayes' rule, we also identify $F_1(w_1|h, d)$, $f_{2|1}(w_2|w_1, h, d)$.*

The detailed proof of the identification theorem is in Appendix A. We here sketch the proof to emphasize the roles of Assumptions 2, 3 and 4. The identification argument is similar, yet not identical, to the ones in the seminal papers of Hu (2008); Hu and Schennach (2008); Kasahara and Shimotsu (2009b); Hu and Shum (2012). We also refer the reader to Hu's (2017) survey and to his forthcoming book on the econometrics of unobservables.

Sketch of the algebraic method of proof. To see how these three assumptions are crucial for identification, at least with the method of proof used in this paper, let us examine the simple case of just two wages, no training, and no wage dynamics. The likelihood of the joint event $\{w_{i1} \leq w_1, w_{i2} \leq w_2, z_i = z\}$ is

$$p(z, w_1, w_2) = \sum_h \pi(h, z)\, F(w_1|h)\, F(w_2|h).$$

Hence we want to identify a discrete mixture with an additional measurement of the latent variable h, which is z. This additional measurement z can be as simple as a binary variable. In Hu (2017)'s classification, we have a 2.1-measurement model. What is important is that z be correlated with h but not with wages given h. Let us consider a grid of wages, and let us store the discretized function $F(w|h)$ in a matrix $G = [F(w|h)]_{w,h}$ where wage points index rows and latent types index columns (as in Bonhomme et al., 2019). The first row of G is made of ones if the first point of the grid is the maximal wage. Next, let $P(z) = [p(z, w_1, w_2)]_{w_1,w_2}$ store the likelihood values in a matrix where w_1 indexes rows and w_2 indexes columns. Finally, let $D(z) = \text{diag}\,[\pi(1, z), ..., \pi(H, z)]$ be a diagonal matrix with the latent type probabilities $\pi(h, z)$ along the diagonal. Then, $P(z) = GD(z)G^\top$.

This looks like an eigendecomposition formula, except that matrix G is not orthogonal. This is why we need two matrices $P(0), P(1)$, using a classic algebraic trick for identifying latent structure models such as finite mixtures and independent component analysis (see for example Cardoso, 1989). Assumptions 2 and 3 guarantee that G and $D(0)$ are full rank. So $P(0)$ has rank H and the number of types is identified. To simplify, let us assume that G is a square matrix (the number of wages on the grid is reduced to H). Let us consider the spectral decomposition: $P(0) = Q\Lambda Q^\top$, with Q symmetric and orthogonal. We can write

$$Q^\top P(0) Q\Lambda^{-1} = Q^\top G D(0) G^\top Q\Lambda^{-1} = I_H,$$

where I_H is the identity matrix. Define $W = Q^\top G$. Then, $D(0)G^\top Q\Lambda^{-1} = W^{-1}$. It also

follows that

$$\begin{aligned} Q^\top P(1) Q \Lambda^{-1} &= Q^\top G D(1) G^\top Q \Lambda^{-1} \\ &= Q^\top G D(1) D(0)^{-1} D(0) G^\top Q \Lambda^{-1} \\ &= W D(1) D(0)^{-1} W^{-1}. \end{aligned}$$

This last expression gives the eigendecomposition of the matrix $Q^\top P(1) Q \Lambda^{-1}$. Its eigenvalues are the elements of the diagonal matrix $D(1)D(0)^{-1} = \text{diag}\left[\frac{\pi(h,1)}{\pi(h,0)}\right]$.

By Assumption 4, the ratios $\frac{\pi(h,1)}{\pi(h,0)}$ are all distinct, meaning that the eigenvalues of matrix $Q^\top P(1) Q \Lambda^{-1}$ are simple. It follows that the eigenvectors in W are identified up to scale. And W contains the information on G, the mixture components. The unknown scale of the columns of W is identified by the property that the columns of G are bounded by one. If the eigenvalues are not simple, then many choices are possible for the basis of the eigenspaces, and only some linear combinations of the mixture components $F(w|h)$ will be identifiable.

To sum up, the role of the instrument is to create two observable matrices $P(1)$ and $P(0)$ with the same algebraic structure. One is used to standardize the other (this is called "whitening" in the Independent Component Analysis literature), by a sort of matrix division operation that gives to the ratios $\frac{\pi(h,1)}{\pi(h,0)}$ the interpretation of eigenvalues. Assumption 4 is also a condition for the point identification of the mixture cdfs $F(w|h)$.

In the detailed proof in Appendix A, we generalize this procedure to the case of Markovian wages (building on Hu and Shum, 2012). This proves identification given treatment d_i and first post-treatment wage w_{i2}. However, how do we know that one group that we have labelled 1 for one particular value of (d, w_2) is the same as the group we have labelled 1 for another value? Assumption 4 allows to align groups across different wage values w_2. The odds ratios $\frac{\pi(h,1,d)}{\pi(h,0,d)}$ being independent of wages w_2, and all different by assumption, they also allow to identify a common labelling of the latent groups over all wages $w_2 \in \mathcal{W}_2(d)$. Across different treatments d, wage predetermination fulfill the same role. □

We are now equipped with a nonparametric identification result and we can safely develop a method to estimate our model. Our estimation method is described in sections 3.3 and 3.4 below. Although the identification proof is constructive, it leads to complicated estimating equations that do not use all the available information. This is why we prefer, for estimation, to use maximum likelihood and a parametric version of the model.[7]

[7] Our parametric version could be made arbitrarily flexible, but the data that we use would not support the estimation of a complicated specification with a large number of parameters. Estimating a parametric model after showing nonparametric identification is standard. See for example Cunha et al. (2010); Bonhomme et al. (2019).

2.2 Treatment effects and usual estimators

Before turning to the estimation procedure and to our empirical application, we discuss the definition of policy-relevant parameters in our framework. We mainly compare the treatment effects with the usual estimators of applied econometrics, such as OLS and IV estimators.

Let $y(0)$ and $y(1)$ denote the *counterfactual outcomes*. In our application, it can be the wages in period $t = 2$ or $t = 3$ of untrained and trained workers, or the wage changes between $t = 2, 3$ and $t = 1$ given training. Hence, our discussion will encompass both static and dynamic experiments (yet not staggered treatments). Note also that, in our setup, counterfactual outcomes $y(0)$ and $y(1)$ satisfy the conditional independence assumption:

$$y(0), y(1) \perp\!\!\!\perp d, z \,|\, h. \tag{3}$$

The difficulty here is that the conditioning variable h is not observed.

Define the observed outcome $y = d\, y(1) + (1 - d)\, y(0)$. We now define and derive the *Average Treatment Effect* (ATE) and the *Average Treatment Effect on the Treated* (ATT). Then we consider OLS and IV estimators.

ATE. We define a *conditional* Average Treatment Effect given type h as follows,

$$ATE(h) = \mathrm{E}[y(1) - y(0)|h] = \mu(h, 1) - \mu(h, 0),$$

where $\mu(h, d) = \mathrm{E}[y(d)|h]$. The unconditional ATE is simply the average over types $h = 1, \ldots, H$ of the conditional ATEs, that is,

$$ATE = \sum_h \pi(h)\, ATE(h), \tag{4}$$

where $\pi(h) = \sum_{z,d} \pi(h, z, d)$ is the population share of type-h workers.

ATT. Under the above conditional independence assumption,

$$ATT(h) = \mathrm{E}[y(1) - y(0)|h, d = 1] = ATE(h).$$

The ATT is thus the average value of the conditional treatment effect $ATE(h)$ over the treated individuals:

$$ATT = \mathrm{E}[y(1) - y(0)|d = 1] = \sum_h \pi(h|d = 1) ATE(h), \tag{5}$$

with $\pi(h|d) = \sum_z \pi(h, z|d)$ and for $d = 0, 1$,

$$\pi(h, z|d) = \frac{\pi(h, z, d)}{\sum_{h,z} \pi(h, z, d)}.$$

OLS and DiD. Now, we study the OLS estimator of the impact of treatment on the outcome. The *difference-in-difference* (DiD) estimator is the OLS estimator when the outcomes $y(1)$, $y(0)$ are defined as wage changes between before and after the treatment's application.

We have

$$\begin{aligned} b_{OLS} &= \frac{\text{Cov}(y, d)}{\text{Var}(d)} = \text{E}[y(1)|d = 1] - \text{E}[y(0)|d = 0] \\ &= \sum_h \pi(h|d = 1)\, \mu(h, 1) - \sum_h \pi(h|d = 0)\, \mu(h, 0) \\ &= ATT + B_{OLS}, \end{aligned}$$

where B_{OLS} is the bias, defined as

$$B_{OLS} = \sum_h [\pi(h|d = 1) - \pi(h|d = 0)]\, \mu(h, 0). \tag{6}$$

Hence, the OLS estimator is an unbiased estimator of ATT ($B_{OLS} = 0$) if

1. $\pi(h|d = 1) = \pi(h|d = 0)$ for all types h; or
2. $\mu(h, 0) = \mu(1, 0)$ for all h.

These restrictions will not hold in general as we expect neither the decision to treat, nor the outcome levels to be independent of individual types. However, with outcomes defined as wage changes between periods before and after training, assumption 2 is the usual common trend assumption in DiD setups: the expected change in the outcome, before and after treatment, is independent of the group. Hence, for levels, we shall refer to B_{OLS} simply as the "heterogeneity" bias. For differences, we will call B_{OLS} the "heterogeneous trend" bias.

Lastly, the sign of the bias is unknown *a priori*. However, imagine that good types, with higher pre-treatment wages (and wage growth), also have a higher probability of benefiting from training. Then, we expect the OLS (or DiD) estimator to be biased upward vis-a-vis the ATT. One can find a similar discussion in Carneiro et al. (2011).

IV and LATE. Finally, the IV estimator of the regression of y on d, using z as an instrument can be expressed as follows,

$$b_{IV} = \frac{\text{Cov}(y, z)}{\text{Cov}(d, z)} = \frac{\text{E}(y|z = 1) - \text{E}(y|z = 0)}{\text{E}(d|z = 1) - \text{E}(d|z = 0)}.$$

First, the denominator of b_{IV} is trivially

$$\mathrm{E}(d|z=1) - \mathrm{E}(d|z=0) = \sum_h \left[\pi(h, d=1|z=1) - \pi(h, d=1|z=0)\right].$$

Second, the numerator can be factored as

$$\begin{aligned}
\mathrm{E}(y|z=1) - \mathrm{E}(y|z=0) &= \sum_h \left[\pi(h, d=1|z=1)\,\mu(h,1) + \pi(h, d=0|z=1)\,\mu(h,0)\right] \\
&\quad - \sum_h \left[\pi(h, d=1|z=0)\,\mu(h,1) + \pi(h, d=0|z=0)\,\mu(h,0)\right] \\
&= \sum_h \left[\pi(h, d=1|z=1) - \pi(h, d=1|z=0)\right]\left[\mu(h,1) - \mu(h,0)\right] \\
&\quad + \sum_h \left[\pi(h|z=1) - \pi(h|z=0)\right]\mu(h,0),
\end{aligned}$$

making use of

$$\pi(h, d|z) = \frac{\pi(h,z,d)}{\sum_{h,d}\pi(h,z,d)} \quad \text{and} \quad \pi(h|z) = \sum_d \pi(h, d|z).$$

Hence,

$$b_{IV} = LATE + B_{IV},$$

where we define

$$LATE = \frac{\sum_h \left[\pi(h, d=1|z=1) - \pi(h, d=1|z=0)\right] ATE(h)}{\sum_h \left[\pi(h, d=1|z=1) - \pi(h, d=1|z=0)\right]} \tag{7}$$

and

$$B_{IV} = \frac{\sum_h \left[\pi(h|z=1) - \pi(h|z=0)\right]\mu(h,0)}{\sum_h \left[\pi(h, d=1|z=1) - \pi(h, d=1|z=0)\right]}. \tag{8}$$

LATE is a weighted average of conditional ATEs given type. This average is informative if the weights are uniformly positive or negative, that is, if monotonicity holds (Imbens and Angrist, 1994):

$$\pi(h, d=1|z=1) \geq \pi(h, d=1|z=0).$$

In our setup, it makes sense to think that the probability of training increases if the employer informs its workers about training possibilities. However, our estimator is more generally applicable as we do not need to assume monotonicity in the treatment probability. As in de Chaisemartin and d'Haultfoeuille (2020)'s application to difference-in-difference, we can check whether all weights are of the same sign or not.

The IV estimator is an unbiased estimator of LATE ($B_{IV} = 0$) if

1. $\pi(h|z=1) = \pi(h|z=0)$ for all types h; or

2. $\mu(h,0) = \mu(1,0)$ for all h.

The second restriction has already been discussed in the case of OLS. The first restriction is also similar, although it now links heterogeneity h to the instrument z instead of the treatment d. In our application, the instrument is determined at the firm level. So, it may be correlated with worker types either because of matching — good firm types matching with good worker types — or if employers themselves inform workers about training possibilities in a selective way. In many usual LATE setups, the instrument is not local (a policy designed at some regional level, for example). In which case, the first restriction is also more likely to hold (that is, if individuals do not move in response to the policy). In randomized setups, z is the intention to treat, the random assignment to treatment and is by construction exogenous. Then, treated individuals may comply ($d = 1$) or not ($d = 0$) with the assignment to treat (*e.g.*, Abadie et al., 2002).

Conclusion. Our setup, therefore, offers two main advantages: 1) It allows one to identify average treatment effects (and more generally their distribution across latent types) in situations where counterfactual outcomes are heterogeneous. In a difference-in-difference setup, this means that identification does not rest on the common trend assumption. 2) Identification is complete, meaning that all parameters of the structural model are nonparametrically identified. This allows one to identify not only the conditional treatment effects given types, but also the joint distribution of treatment and types. Hence, the weights of marginal treatment effects in OLS and IV estimators can be separately identified, with no need for such assumptions as constant-sign or monotonicity.

3 Application: the wage returns to training

3.1 The data

We use survey data collected between 2013 and 2015 by Céreq,[8] as part of the DEFIS survey.[9] The survey sampled 4,529 firms with three employees or more from all sectors but agriculture in 2013, and 16,126 workers were subsequently drawn from these firms' employees.[10] The main objective of the survey was to document the use of formal or non-formal adult education by employees, and the effect of this form of learning on work outcomes. Several waves of interviews were conducted. We use the first wave in this paper, in which employees were interviewed between June and October 2015 about any training sessions that they participated in between January 2014 and the time of the interview.

[8] *Centre d'études et de recherches sur les qualifications* (a French public institution).

[9] *Dispositif d'enquêtes sur les formations et itinéraires des salariés.*

[10] The employees were sampled among the sampled firms' employees, provided that they were employed by their firm in December 2013. The latter sampling is stratified to provide a representative sample of workers

This was done through retrospective questions (such as "Did you hold a full-time or a part-time contract in firm X in the fall of 2013?", or "Since January 2014, did you take part in a training program?").

The responses to the employer survey (in December 2014) and the worker survey (in 2015) are matched with wage data obtained from tax registers, reported by employers to the tax authorities (*Déclarations annuelles de salaires*, DADS) for the ongoing employment spells in December 2013, December 2014, and December 2015.[11] Our definition of the wage is the total earnings paid to the worker by the employer in December 2013, 2014, and 2015, net of payroll taxes (but not net of income tax) and divided by the total number of hours worked in that employment in the whole years of 2013, 2014, and 2015. Nearly 80% (12,597/16,126) of workers reported that they were employed by the same firm as in 2013 at the time of the interview in 2015. Greater fractions (89.2% = 12,100/13,562 in 2014 and 85.3% = 11,103/13,014 in 2015) of the wages recorded for 2014 and 2015 were paid by the same employer who paid the wage recorded in 2013. Therefore, a large majority of workers in our data did not move during our period of analysis so we will abstract from worker mobility in this paper.

To give a first overview of the factors affecting the selection into training, we start with a simple comparison of employees who reported at least one training session in 2014 or 2015 with employees who did not declare any training. Among the 16,126 employees surveyed in 2015, 6,349 individuals (39.3%) declared at least one training session, with a majority of them declaring only one session.[12] Table 1 presents the average characteristics of trained and untrained workers in terms of demographics, education, occupation, job, and firm characteristics, before any training (situation in the fall of 2013). Statistics are presented both for the overall sample (the two left-hand columns) and the analysis sample (the two right-hand columns). The analysis sample excludes some individuals with extreme wage observations and more importantly, includes only "stayers" — workers who are observed in the same firm in 2013 and in 2015.

All variables in rows are binary, except the age and hourly wage (in logs). Table 1 suggests that on average, workers who trained between January 2014 and the time of the first interview (between June and October 2015) are more likely to be French, male, living as a couple, and to have children (even controlling for age) compared to workers who did not train. They also tend to be more educated, most of them having post-secondary degrees. They occupy more skilled jobs, they have higher salaries, and they are more likely to hold full-time and permanent contracts. They are also more likely to receive information on training (our instrument). Using the employer survey, we also find that trained workers are on average in bigger firms, that are more likely to have human resource

[11] More precisely, the last employment spells of the years 2013, 2014 and 2015, which ends at the end of December for 83% of the workers in 2013, 78% in 2014 and 76% in 2015.

[12] Among the 6,349 employees who received training, 61% declared one session, 26% declared two, 9% declared 3, and less than 4% declared more than 3.

Table 1: Comparison of trained and untrained workers by baseline characteristics

	All		Stayers	
	Trained	Untrained	Trained	Untrained
Demographics:				
Age (modal group)	40-44	45-49	40-44	45-49
Male	70.7	67.3	74.1	72.9
French	97.0	94.1	98.1	95.8
In couple	74.8	68.4	78.6	74.0
Has children	57.4	49.0	63.1	55.9
Disability	7.2	12.5	6.7	10.1
Previous health problem	3.4	5.7	2.6	4.1
Education:				
Less than high school diploma	28.3	46.1	29.4	46.5
High school diploma	18.5	18.6	18.3	18.1
Trade or vocational degree	20.7	14.9	21.9	16.7
Bachelor's degree	7.9	5.5	6.9	4.7
Master's degree or more	23.9	13.8	23.0	13.1
Occupation:				
Unskilled blue collar	5.9	9.6	5.2	9.0
Skilled worker, technician	18.5	26.2	18.8	26.8
Office worker, public sector employee	21.2	27.9	17.5	24.3
Foreman/Supervisor	13.7	9.9	15.0	11.2
Technician, draftsman, salesman	9.3	6.5	10.0	7.4
Engineer, manager	29.5	15.7	32.7	18.9
Job characteristics:				
Log(hourly wage), 2013 (w_1)	2.7	2.5	2.8	2.6
Log(hourly wage), 2014 (w_2)	2.8	2.6	2.8	2.6
Log(hourly wage), 2015 (w_3)	2.8	2.6	2.8	2.6
Permanent contract	90.0	83.3	98.5	98.3
Full time contract	88.7	80.1	95.9	93.9
Information on training (z)	78.8	62.8	81.7	68.5
Firm characteristics:				
3 to 49 employees	24.0	39.1	21.1	38.2
50 to 249 employees	20.5	21.8	21.4	23.5
250 to 499 employees	9.1	7.2	9.7	7.9
500 to 999 employees	8.6	6.5	8.5	6.8
1000 to 1999 employees	7.4	6.2	7.2	5.4
More than 2000 employees	30.4	19.1	32.1	18.3
Has HR department	89.6	81.5	91.5	81.9
Has individual incentive strategy	72.4	60.0	74.4	61.8
Has collective incentive strategy	78.4	64.5	82.2	68.6
Outsources part of activity	40.6	34.8	41.9	36.5
Number of observations	6343	9783	3467	4066

Notes: "All" refers to the whole sample and "Stayers" refers to the sub-sample of workers who remain employed in the same firm all three years. For all binary variables, the mean is given as a percentage. The bottom row gives the number of workers for all variables except log(hourly wage), where 59 observations are missing wages in 2013, and approx. 3,000 in 2014 and 2015.

staff. Overall, more advantaged workers are more likely to get training. The two samples are generally similar across observable dimensions, with notable differences being that individuals in our analysis are more likely to be full-time and hold a permanent contract.

In the next section, we present the results from estimating, by OLS and IV, a system of equations that resembles the model presented in section 2 for our application. This allows us to compare the results using our method to those obtained using standard approaches, the theoretical analysis of Subsection 2.2 having demonstrated the potential biases on OLS and IV estimators.

3.2 Preliminary analysis

We start by estimating the wage equation,

$$w_{it} = \alpha_t + \beta_t d_i + x_i \theta_t + v_{it}, \tag{9}$$

where w_{it} are log-wages at the end of 2013 ($t = 1$), 2014 ($t = 2$), and 2015 ($t = 3$); d_i is an indicator for training between January 2014 and December 2015; and $x_i\theta_t$ is a combination of control variables (as observed in 2013).[13] This equation is first estimated by OLS for each year separately, and then by 2SLS, instrumenting d_i by z_i, the *information on training* mentioned above. The estimations are done with and without controls. The DiD estimate of the effect of training in 2014 and 2015 is obtained as $\Delta\beta_2 = \beta_2 - \beta_1$ and $\Delta\beta_3 = \beta_3 - \beta_1$.

The results are reported in Table 3. The OLS results suggest very small effects of training (differences in the β's around 0.2-0.5% with controls) and the effect of training on pre-treatment wages remains significant even after adding many controls to the estimation. After instrumenting the training variable, we see both stronger effects of around 4%, and the effect of training on 2013 wages stops being significant when controls are included in the regressions. Note that standard errors jump by one order of magnitude, pointing at a certain weakness of the instrument.

These results suggest the existence of a causal link between wages and training of around 4%, which is non-negligible. We now use our model in order to check whether there is any reason to doubt that the IV estimation delivers an unbiased estimate of the causal effect of training on wages.

[13] For controls, we use: gender, age brackets, married, handicapped, having health problems, open-ended contract, full-time contract, socioeconomic status, firm size brackets, existence of an HR department, existence of wage incentives for performance (individual and collective), whether the firm outsources activities. See Table 1 for summary statistics.

Table 3: Static estimation of wage regressions with training

	OLS		2SLS	
	Without controls	With controls	Without controls	With controls
Log-wage levels				
2013	0.158	0.038	0.179	0.057
	(0.009)	(0.006)	(0.060)	(0.053)
2014	0.166	0.040	0.219	0.098
	(0.009)	(0.006)	(0.061)	(0.053)
2015	0.169	0.043	0.216	0.093
	(0.009)	(0.006)	(0.062)	(0.054)
Log-wage changes				
2014	0.007	0.002	0.040	0.041
	(0.003)	(0.003)	(0.019)	(0.025)
2015	0.011	0.005	0.037	0.036
	(0.003)	(0.004)	(0.022)	(0.030)
Nb of workers	7,533	7,533	7,533	7,533

3.3 Parametric specification

In practice, we specify a parametric version of the model and we use maximum likelihood for estimation.

We assume that log-wages are normal conditional on type and training, and first-order autoregressive with autocorrelation coefficient ρ. More precisely, we postulate that

$$w_1 = \mu_1(h) + u_1, \quad \text{where} \quad u_1 \sim \mathcal{N}(0, \sigma_1^2(h)),$$

and for $t = 2, 3$,

$$w_t = \mu_t(h, d) + u_t, \quad \text{where} \quad u_t \sim \mathcal{N}(\rho u_{t-1}, \sigma_t^2(h, d)).$$

Then, with $\varphi(u) = (2\pi)^{-1/2} e^{-u^2/2}$, we have,

$$f_1(w_1|h) = \frac{1}{\sigma_1(h)} \varphi\left(\frac{w_1 - \mu_1(h)}{\sigma_1(h)}\right),$$

and

$$f_{2|1}(w_2|w_1, h, d) = \frac{1}{\sigma_2(h, d)} \varphi\left(\frac{w_2 - \mu_2(h, d) - \rho\left[w_1 - \mu_1(h)\right]}{\sigma_2(h, d)}\right),$$

$$f_{3|2}(w_3|w_2, h, d) = \frac{1}{\sigma_3(h, d)} \varphi\left(\frac{w_3 - \mu_3(h, d) - \rho\left[w_2 - \mu_2(h, d)\right]}{\sigma_3(h, d)}\right).$$

The model is flexible at first and second order as long as parameters μ_t, σ_t are left unrestricted. A more flexible distribution than the normal could be used for the distribution of innovation errors, but, as we shall see, the link between wages and training is tiny. Thus, there is little data to infer higher-order moments.

Probabilities $\pi(h, z, d)$ are left unrestricted.

The data for each individual i is the array $x_i = (w_{i1}, w_{i2}, w_{i3}, z_i, d_i)$. The parameters of the model are denoted $\beta = (\mu, \pi, \rho, \sigma)$. The complete likelihood of individual i's observations x_i and any type h is

$$\begin{aligned} \ell_{ih}(\beta) &\equiv \ell(x_i, h, \beta) \\ &= \pi(h, z_i, d_i)\, f_1(w_{i1}|h, \beta)\, f_{2|1}(w_{i2}|w_{1i}, h, d_i, \beta)\, f_{3|2}(w_{i3}|w_{i2}, h, d_i, \beta). \end{aligned} \tag{10}$$

The individual likelihood is $\ell_i(\beta) = \sum_h \ell_{ih}(\beta)$. The sample likelihood is the product of individual likelihoods, $L(\beta) = \prod_i \ell_i(\beta)$.

3.4 Types and likelihood maximisation

We found that a two-stage approach to estimation worked best in our application. In the first stage, we classify workers into types based solely on their wages, abstracting from training and training information. We then perform a second round of classification within each group from the first stage, now allowing wages to depend on training. Within each stage, the EM algorithm is used to estimate the discrete mixture, and groups are labelled by increasing values of mean wages in 2013, i.e. by μ_1. Specifically, we now assume that each type h is a pair $h = (k, g)$, where $k \in \{1, \ldots, K\}$ is the first-stage type component (depending only on wages) and $g \in \{1, \ldots, G\}$ is a second-stage type component (depending on training and wages). It follows that the total number of discrete groups is $H = GK$.[14]

We started by estimating the full model with unrestricted types in a single step. However, most of the latent classification was used to fit the overall distribution of wages, and little heterogeneity was spared to fit different relationships between wages and training. In particular, it was not possible to avoid wages in 2013 varying with training in 2014 within each estimated group. By first estimating a discrete mixture of wages and then re-estimating a discrete mixture of wages *and* training, given the first classification, we increase our chances of zooming in on the wage-training link. Note that, in principle, this two-stage procedure is used without loss of generality. Indeed, nothing prevents the estimated second-stage mixture from being exactly the same within each of the first-stage groups.

[14] We could have let G depend on k, each first-stage type k determining a different number of second-stage types $G(k)$, but to keep the analysis relatively simple, we keep G constant across k.

The simplified first-stage model is

$$w_1 = \bar{\mu}_1(k) + u_1, \quad u_1 \sim N\left(0, \bar{\sigma}_1^2(k)\right),$$
$$w_t = \bar{\mu}_t(k) + u_t, \quad u_t \sim N\left(\bar{\rho} u_{t-1}, \bar{\sigma}_t^2(k)\right), \quad t = 2, 3,$$

where we use an upper bar to distinguish the first-stage variables and parameters from those of the second stage.

We use a sequential EM-algorithm for the likelihood maximization of both stages (see Appendix B for details). Moreover, we relabel groups k and subgroups g by increasing values of $\overline{\mu}_1(k)$ and $\mu_1(k, g)$ after estimation has converged.

3.5 Estimating the number of types, H

In the identification of our model, one of the key parameters that we showed to be identified was the number of types, H. In our identification strategy, the number of types was simply the rank of the matrix of observed data points, $P(z)$. However, in the alternative method we use to estimate our model, the econometrician fixes H at the start of the procedure. Therefore, if we want to avoid selecting the number of types arbitrarily, we need a method to estimate (or "choose") H. This problem has been well-studied theoretically in the computer science literature, although practical methods are rare, especially in situations where the correct model is not in the set of considered models (Fraley and Raftery, 1998).

As the number of groups increases, so does the number of parameters. Hence the fit of the model is monotonically increasing in the number of groups. Therefore, the "elbow-method" has been widely used, which involves looking for "elbows" in some objective function, i.e. where it starts to increase less steeply. We use a range of (penalized) likelihood functions as the objective function in our analysis, including the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the integrated conditional likelihood (ICL) proposed by Biernacki et al. (2000). The ICL criterion was proposed to counter the tendency of BIC to overestimate the number of groups by penalizing the likelihood when groups are not well separated.[15]

3.6 Bootstrap

Standard calculations of parameter standard errors do not incorporate the random nature of the estimated classification (even if it should be negligible asymptotically). We therefore bootstrap standard errors by resampling and reestimating many times the whole

[15] As pointed out by Biernacki *et al.*, the BIC is a reliable approximation of the integrated likelihood if the estimated parameters are well within their domain. This is not the case if the estimated K is greater than the true K^0, as $K - K^0$ shares should be equal to 0.

procedure. This is computationally intensive as we use 500 replicated samples, with replacement, from the original sample. Specifically, we use the weighted-likelihood bootstrap. O'Hagan et al. (2019) show that it provides a robust solution in our setting. Standard bootstrap may generate unstable results if re-sampling causes certain types to be under-represented or even to disappear. The weighted version draws non-zero weights for each observation from a Dirichlet distribution to ensure that no observations are completely dropped in any bootstrapped sample (Newton and Raftery, 1994). The weights λ_i are such that they sum to the size of the full sample, that is, $\sum_i \lambda_i = N$. We use the original, full-sample estimates as initial values for the algorithm at the beginning of each re-estimation. Confidence intervals can then be estimated by selecting the corresponding percentiles of the bootstrapped parameter estimates, i.e. the 5th and 95th percentiles for a 90% confidence interval.

4 Results

4.1 Choosing the number of types (K and G)

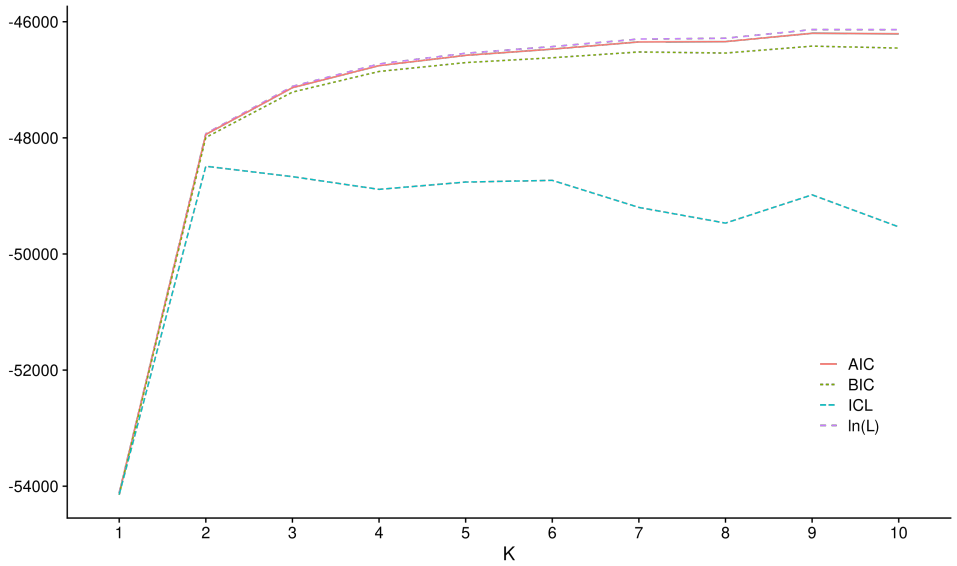
Our estimation strategy requires the econometrician to choose the number of types in both stages, K and G. Figure 1 presents some of the criteria we use to choose the number of types for the remainder of our analysis. In Figure 1a, the different broken lines show how total likelihood ($\ln L$) and penalized-likelihood criteria evolve with K. The first two penalized-likelihood criteria are the well-known Akaike and Bayesian information criteria (respectively, AIC and BIC). We are looking for "elbows", that is, values of K where the marginal gain in likelihood for an additional type is noticeably less than it is for $K-1$. There is a clear elbow at $K=3$ or $K=4$ for AIC and BIC. The ICL criterion is more or less steadily decreasing for all K.

In Figure 1b we study what happens to the sizes and means of the groups as we increase K. The groups all appear distinct when compared by their means, but very small groups start to appear for $K=6$ or 7. Notice also that there is only a small fraction of the workers who display different mean wages across periods. There is only one group with clearly different wage means for $K \leq 5$. For $K=6$ or 7 we see more than one group with different wage means, but this looks like a dilution of the only such group appearing when $K=4$ or 5. Combining the evidence from both panels of Figure 1, we choose $K=4$ for the first stage, and we leave to the second stage the task of determining the role of training in generating the observed changes on mean wages over time.

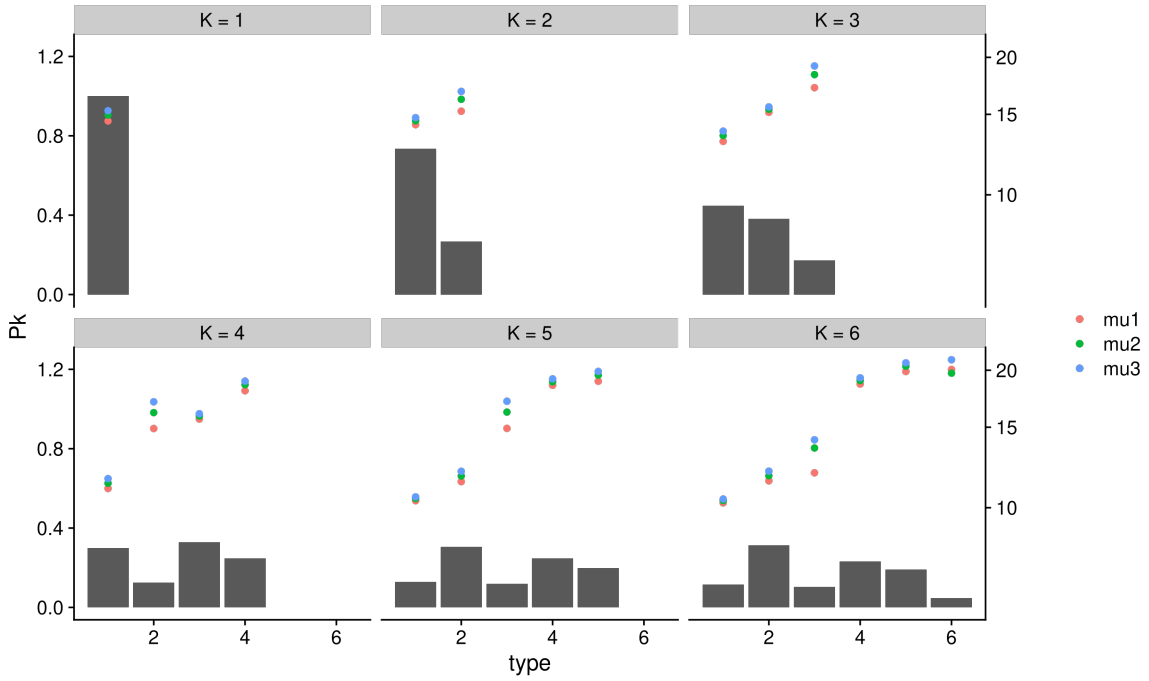
Figure 2 represents the results from the second stage of our estimation procedure. In panel (a), each of the four subpanels shows the likelihood criteria for each one of the four types obtained in the first stage. The likelihood, here, is a weighted sum of individual likelihoods where the weights are the posterior type probabilities estimated from the first

Figure 1: Choosing the number of types K

(a) Likelihood criteria



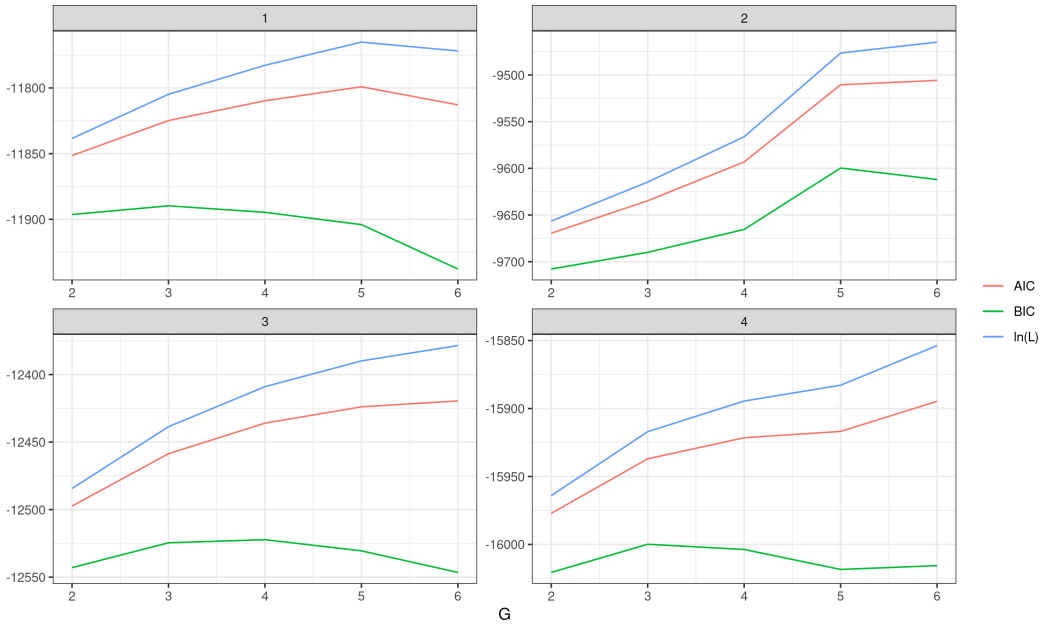
(b) Group sizes (bars) and means (points)



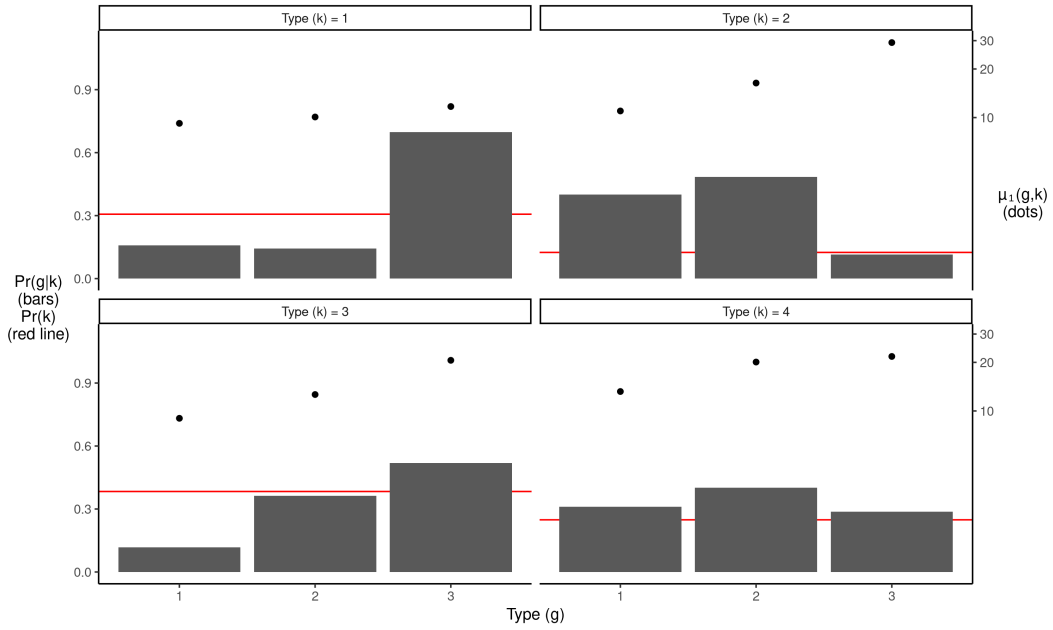
Notes: (a) If M is the number of parameters, N the number of observations, and L the likelihood, $AIC = -\ln L + \frac{1}{2}\ln M$, and $BIC = -\ln L + \ln(N)M$. We plot $-AIC$ and $-BIC$ on the figure. The ICL is an alternative criterion proposed by Biernacki et al. (2000). (b) The bars in Panel (b) are the shares of each group. The colored dots are the levels of estimated mean wages in the three years for which we observe wages.

Figure 2: Choosing G (stage 2)

(a) Likelihood criteria



(b) Group sizes (bars) and means (points), $G = 3$



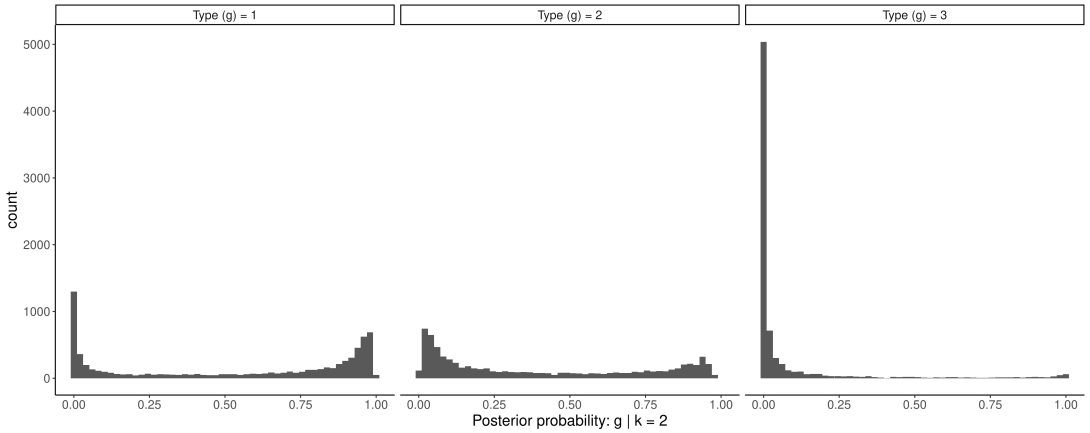
Notes: (a) If M is the number of parameters, N the number of observations, and L the likelihood, $AIC = -\ln L + \frac{1}{2}\ln M$, and $BIC = -\ln L + \ln(N)M$. We plot $-AIC$ and $-BIC$ on the figure. (b) The bars in Panel (b) are the shares of each group. The dots are the estimated mean wages in 2013.

Table 4: Test of assumption 4

$G=$	1	2	3	4
$K=1$	-	0.023	0.071	0.092
2	0.004	0.119	0.107	0.107
3	0.000	0.028	0.085	0.057
4	0.048	0.031	0.019	0.013
5	0.012	0.015	0.007	0.001

Notes: The table shows $\min_{h\neq h',d}\left|\frac{\Pr(z=1,h,d)}{\Pr(z=0,h,d)}-\frac{\Pr(z=1,h',d)}{\Pr(z=0,h',d)}\right|$ for all K and G. This is a criteria to help choose the number of types based on assumption 4.

Figure 3: Conditional posterior type probabilities $p_i(g|k=2)$



Notes: The three panels above show the posterior probabilities of workers of being some type g, conditional on them being type $k=2$. They show the distribution of the second stage types conditional on $k=2$. There are 50 bins in each histogram.

stage (see Appendix B). For all types except $k=2$, the BIC wants $G=3$.[16] When $k=2$, the optimal G is 5, but given that this is the smallest group from stage 1 (represented by the red horizontal lines in panel (b)), and for the sake of simplicity, we choose the same number G of second-stage types within each first-stage type k. To avoid an abundance of numbers and plots, we choose $G=3$ and show results with $G=3$ for all types $k=1,\ldots,4$. We checked that $G=5$ delivers similar conclusions.[17]

In table 4 we present a criterion based on assumption 4 to help select the number of types. Recall that assumption 4 requires that different types have different exposures to the instrument. We can check to what extent this assumption holds for different values of K and G. The values in table 4 are the minimum differences, $\left|\frac{\Pr(z=1,h,d)}{\Pr(z=0,h,d)}-\frac{\Pr(z=1,h',d)}{\Pr(z=0,h',d)}\right|$, across all values of $h\neq h'$ and d. We do not want this minimum difference to be too close to zero for our chosen K and G.

[16]The second-stage ICL actually wants a larger G than the BIC. Given the motivation for the ICL (the BIC can *overestimate* K) we choose the K suggested by the BIC in the second stage.

[17]Our results are stable across a range of values for K and G.

A final criterion to determine the optimal number of groups is whether groups are well differentiated or not — sometimes known as the *entropy* of the classification. This is what the ICL criterion takes into account, and the BIC does not. We can check the strength of the classification by plotting the distribution of posterior group-probabilities, $p_i(h)$.[18] Figure 3 displays these distributions conditional on $k = 2$ and for all $g = 1, 2, 3$, i.e. $p(g|k = 2)$. We see that they are concentrated near zero and one.

4.2 Observed characteristics by type

We did not include controls when estimating the model to avoid the double complication of choosing a functional form for probabilities $\pi(h, z, d)$ and specifying the interaction between observed and unobserved heterogeneity in these probabilities and the wage densities. But we can still study if types can be characterized by some specific values of observed variables. We first assign to each individual a type corresponding to their highest posterior probability, and then study the individuals assigned to each group. The results of this exercise are in Table 5.

Interestingly, although we only use wages in the first stage, and wages and training in the second, the resulting classification does not correspond to any (obvious) classification in terms of other observed characteristics. For example, k is not obviously associated to education, and g is not obviously related to occupation or firm size.

4.3 Parameter estimates (with $K = 4$ and $G = 3$)

Figure 4 displays the probability of being treated conditional on first- and second-stage types, and the value of the instrument, i.e., $\pi(d = 1|k, g, z)$. The error bars indicate bootstrapped, 90% confidence intervals. There are two key features to note. First, right-blue bars are higher than left-red ones. This is evidence of instrument monotonicity, which holds almost perfectly: those who receive information on training are more likely to train across all types (except $(k, g) = (2, 2)$). Second, the bars are generally increasing in both k and g.

In figure 5 we show the different components of the decomposition of assumption 4. We can see that in panels (a) and (b) the bars are generally different sizes, suggesting that assumption 4 likely holds. This remains true when we combine the bars in panel (c), although there are a few that are close.

In Figure 6, we study the correlation between the types and the instrument. As already emphasized, IV is equal to LATE only if the instrument and the type are independent, which would require the red and blue bars to be equal for each combination of k and g, i.e., that $\Pr(k, g|z = 1) = \Pr(k, g|z = 0)$. This assumption seems violated here, and there

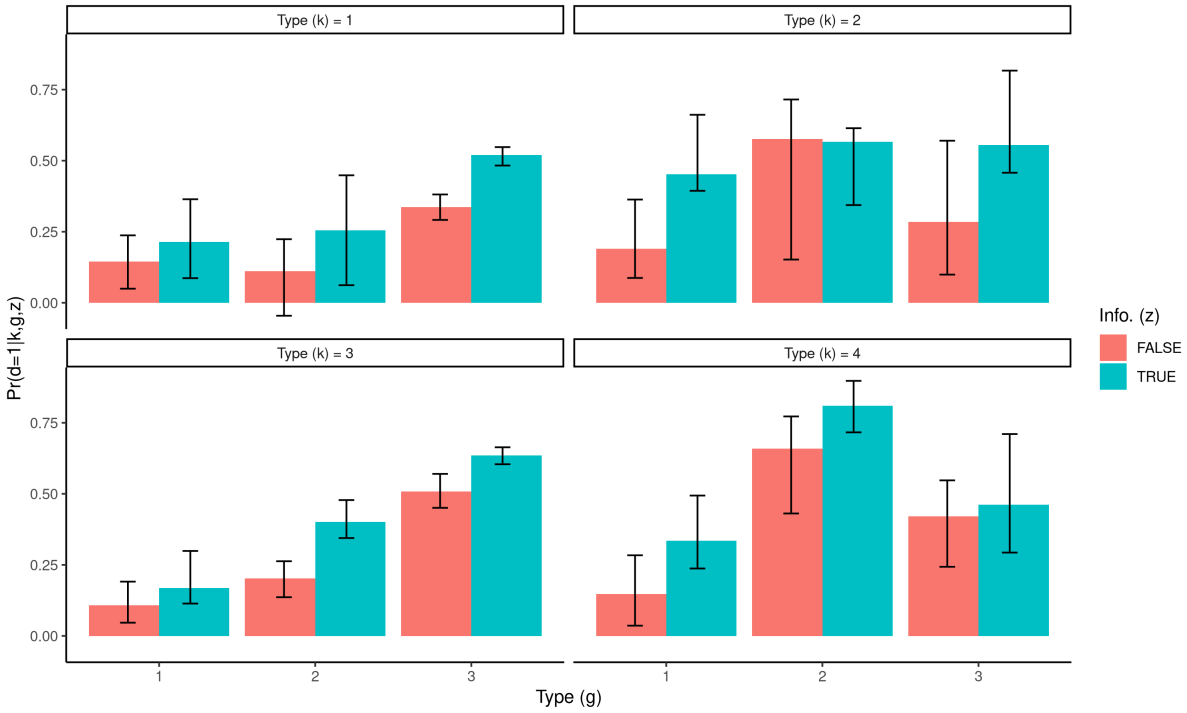
[18] By definition, $p_i(h) = \Pr\{h_i = h \,|\, w_{it}, z_i, d_i\} = \ell_{ih}(\widehat{\beta}) / \sum_h \ell_{ih}(\widehat{\beta})$.

Table 5: Comparing types by baseline characteristics

	$k_i = 1$			$k_i = 2$			$k_i = 3$			$k_i = 4$		
$g_i =$	1	2	3	1	2	3	1	2	3	1	2	3
Demographics:												
Age (modal group)	45	45	40	50	40	55	40	50	50	45	45	50
Male	57.8	66.7	71.1	62.3	68.5	91.1	56.4	73.3	80.7	75.4	81.9	85.2
French	93.1	96.4	96.9	96.9	98.1	94.9	93.9	97.4	97.4	96.2	97.6	97.7
In couple	61.7	66.7	73.4	73.0	78.1	87.3	64.6	75.3	82.2	77.1	82.2	84.4
Has children	44.9	52.9	57.1	53.5	62.1	63.3	48.2	57.4	66.0	57.3	67.2	64.3
Disability	13.8	12.4	9.90	21.1	7.73	5.06	9.2	8.92	4.50	12.2	3.29	5.28
Previous health issue	3.29	5.33	3.86	6.92	2.40	1.27	6.78	4.14	1.69	4.06	1.90	1.51
Education:												
Less than HS diploma	59.6	63.6	46.5	54.7	22.4	13.9	61.5	50.5	18.1	47.3	20.3	21.4
HS diploma	24.9	19.1	22.4	16.0	16.0	8.86	23.0	19.1	13.6	17.9	15.8	12.3
Trade / voc. degree	9.88	10.7	18.3	12.9	22.4	15.2	7.99	17.9	26.0	17.4	23.9	20.9
Bachelor's degree	2.99	3.11	6.04	5.03	8.00	8.86	3.39	6.16	6.04	4.53	6.58	5.53
Master's degree +	2.10	2.22	5.93	10.10	29.9	53.2	2.66	5.98	35.7	12.4	32.9	39.2
Occupation:												
Unskilled blue collar	13.2	15.1	9.38	8.81	1.60	2.53	15.3	9.02	1.33	7.88	1.52	1.26
Skilled, technician	33.2	40.4	32.8	34.6	10.1	5.06	32.9	31.4	7.96	28.2	11.4	6.28
Office, public sector	43.4	27.6	28.8	26.7	10.9	5.06	42.9	25.7	6.71	23.2	8.99	7.79
Foreman/Supervisor	0.90	3.56	11.1	9.12	12.5	1.27	1.21	12.8	12.0	11.7	13.7	10.6
Technician, sales	2.69	6.67	9.55	6.29	10.1	3.80	2.66	12.0	9.36	7.88	10.1	4.02
Engineer, manager	0.60	2.22	5.64	9.43	53.3	81.0	0.73	7.18	61.5	19.1	53.4	68.8
Job characteristics:												
Log(hourly wage)												
2013 (w_1)	2.19	2.26	2.44	2.35	2.80	3.44	2.18	2.53	3.07	2.60	3.04	3.25
2014 (w_2)	2.21	2.28	2.47	2.45	2.89	3.50	2.19	2.54	3.09	2.66	3.06	3.27
2015 (w_3)	2.23	2.29	2.50	2.45	3.01	3.53	2.20	2.55	3.10	2.67	3.08	3.29
Permanent contract	97.6	98.2	97.8	96.2	98.7	96.2	98.5	99.2	99.3	97.9	98.6	98.2
Full time contract	87.7	88.9	94.9	93.4	97.9	94.9	86.7	95.5	96.7	95.9	96.8	97.5
Info. on training (z)	64.4	63.1	78.2	75.2	82.1	75.9	61.3	73.0	78.0	74.7	79.9	62.1
Training (d)	12.3	2.22	50.2	36.5	57.6	45.6	12.6	34.9	62.9	17.7	83.7	40.5
Firm characteristics:												
Number of employees												
3 to 49	47.6	48.9	34.0	31.8	23.7	29.1	49.2	37.1	17.0	31.0	17.8	25.6
5 to 249	25.1	28.0	26.3	22.6	19.7	19.0	24.9	19.4	20.6	22.9	19.0	23.4
25 to 499	7.49	6.22	7.13	10.1	10.9	12.7	5.81	8.28	9.36	11.7	9.37	11.6
50 to 999	4.19	3.11	6.50	8.49	8.80	3.80	6.05	8.28	9.06	8.35	8.23	9.05
1000 to 1999	2.69	4.89	5.70	7.86	7.47	3.80	3.63	5.89	7.44	4.30	7.85	8.54
More than 2000	12.9	8.89	20.4	19.2	29.3	31.6	10.4	21.1	36.6	21.7	37.7	21.9
Has HR department	76.9	75.6	85.6	85.5	92.3	88.6	71.7	82.8	91.5	87.1	93.7	90.2
Individual incentives	53.9	49.8	64.5	62.6	75.5	77.2	44.1	63.8	77.1	67.8	81.8	71.6
Collective incentives	57.5	57.3	72.9	69.2	79.5	78.5	53.5	72.9	85.0	76.1	86.2	76.9
Outsources activity	29.3	27.1	35.2	43.7	43.7	40.5	27.4	34.6	47.9	38.2	45.1	44.7
No. of observations	334	225	1740	318	375	79	413	1090	1360	419	790	398

Notes: This table shows the characteristics of each type. We first assign each individual i the type, $h_i = (k_i, g_i)$, corresponding to their largest posterior probability, i.e. $h_i \equiv \arg\max_h p_i(h)$. Then, having assigned individuals to types, we can treat each type as a separate group and produce summary statistics of that type using the variables in our dataset.

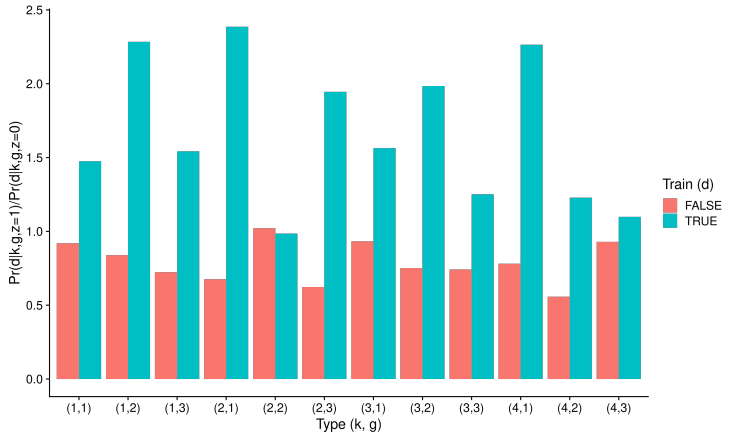
Figure 4: Treatment probability, $\Pr(d = 1|k, g, z)$



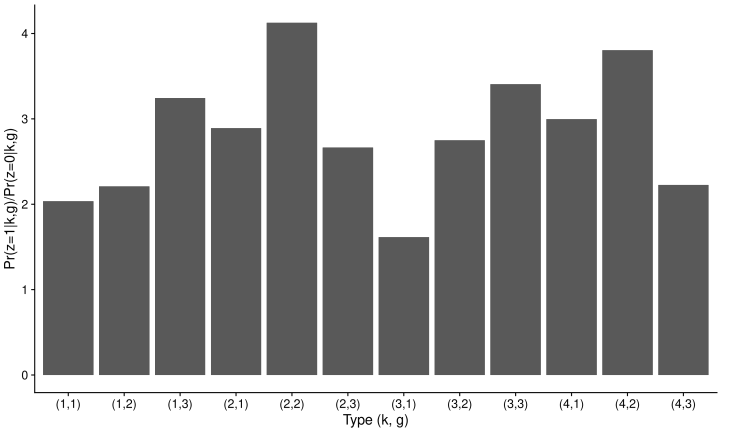
Notes: The bars show the probability of training, conditional on type, and on whether information is provided (intention to treat), i.e. $\Pr(d = 1|k, g, z)$. The red bars correspond to no information ($z = 0$) and the blue ones represent those to whom information was provided ($z = 1$). The second-step types, g, vary along the x-axis, while the panels show different first-stage types, k. The error bars display 90% confidence intervals, obtained by bootstrap.

Figure 5: Test of assumption 4 ($\frac{\Pr(h,z=1,d)}{\Pr(h,z=0,d)} \neq \frac{\Pr(h',z=1,d)}{\Pr(h',z=0,d)}$)

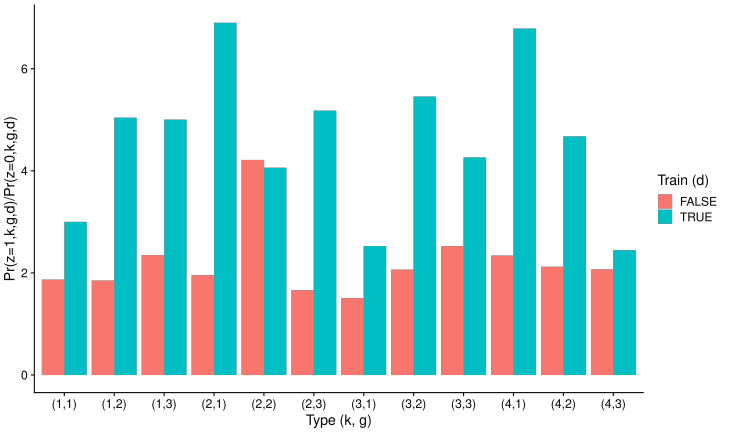
(a) $\frac{\Pr(d|k,g,z=1)}{\Pr(d|k,g,z=0)}$



(b) $\frac{\Pr(z=1|k,g)}{\Pr(z=0|k,g)}$

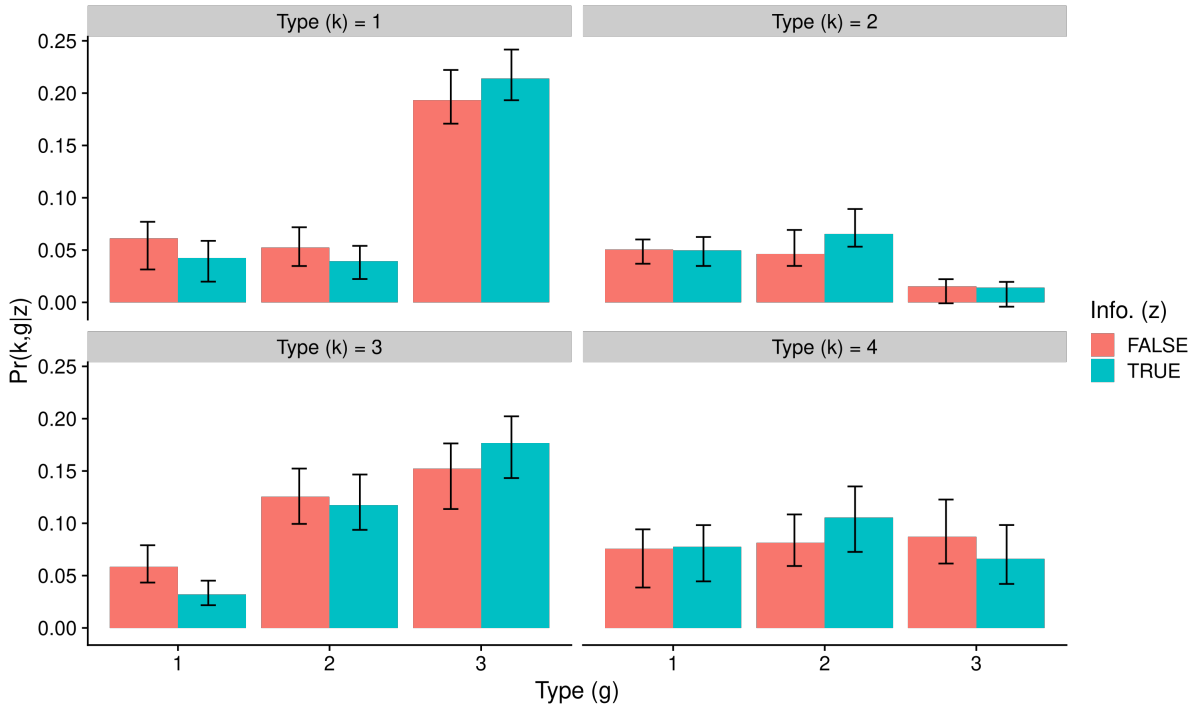


(c) $\frac{\Pr(z=1,k,g,d)}{\Pr(z=0,k,g,d)}$



Notes: This figure shows the components of the decomposition of a key assumption, assumption 4. In panel (a) is the ratio of treatment probabilities for different intentions to treat, with the untrained in red and trained in blue, for each type (x-axis). In panel (b) are the ratios of intention to treat for each type (x-axis). In panel (c) the bars from panels (a) and (b) are combined.

Figure 6: Composition, $\pi(k, g|z)$



Notes: The bars show the probability of being in a given group conditional on intention to treat, i.e. on whether they receive information, with red bars for those who do not receive information ($z = 0$) and blue bars for those who do ($z = 1$). Error bars show 90% confidence intervals, obtained by bootstrap.

seems to be a pattern to the differences in bars. First, subgroups $g = 1, 2$ show similar differences by z, opposite to $g = 3$. Moreover, groups $k = 1, 3$ and $k = 2, 4$ show opposite differences by z, which could be significant as groups $k = 2, 4$ are the ones exhibiting the greatest mean wage variations over time.

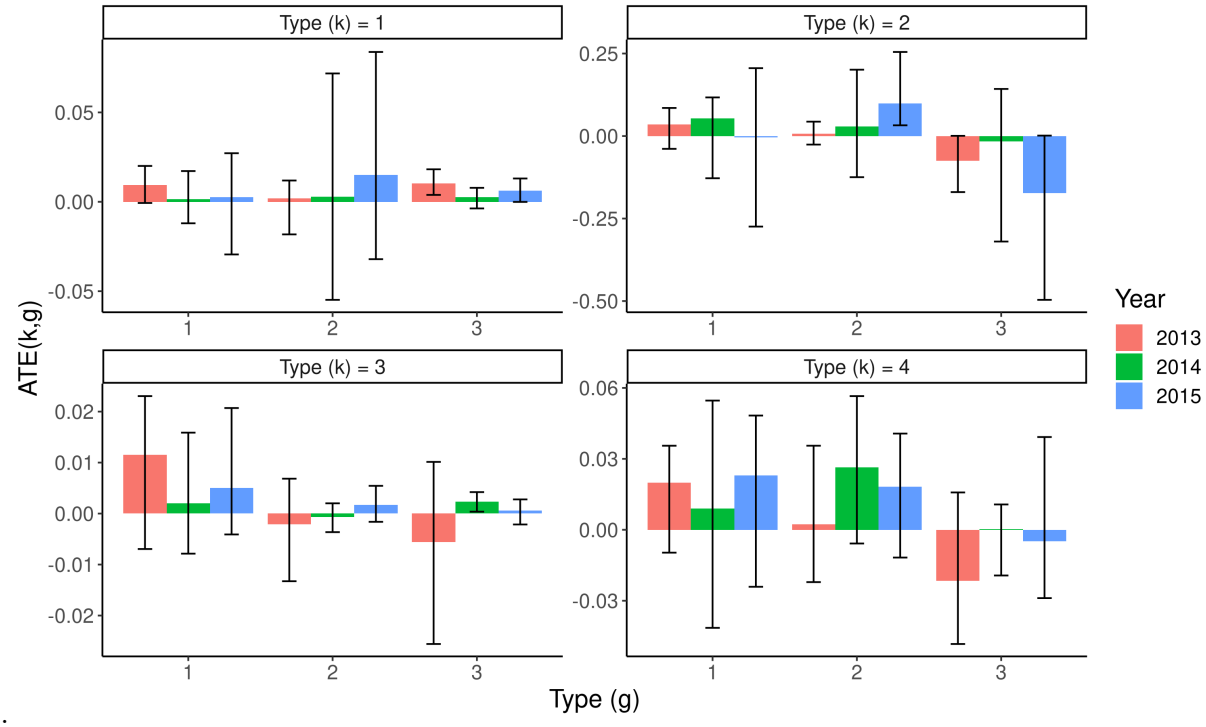
4.4 Treatment effects (with $K = 4$ and $G = 3$)

We now move on to the main objects of our analysis, the treatment effects. Panel (a) of Figure 7 displays treatment effects $ATE(k, g)$, conditional on all types (k, g). Noting that the y-axes differ between cells, we see substantial heterogeneity in treatment effects across types. The effects estimated for $k = 2$ are five times larger than for the rest of the k-types. However, none of these conditional ATEs is very precisely estimated.

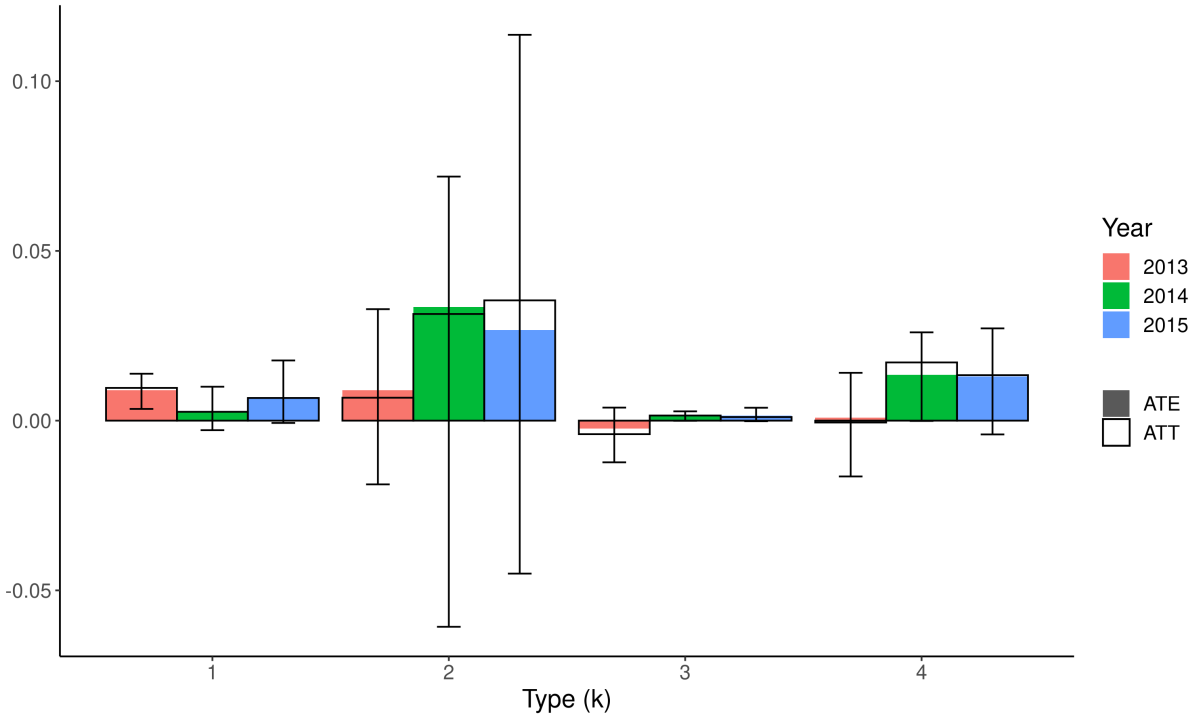
Note that we calculate an empirical wage mean in 2013 for the trained and the untrained, and corresponding pseudo-ATEs, although we estimated mean wages in 2013, $\mu_1(k, g)$, as independent of d. We did this calculation to check this independence assumption *ex post*. Unfortunately, the 2013 pre-treatment effects are of similar size to (or even greater than) 2014 and 2015 post-treatment effects for some types. Together with the large bootstrap standard errors, this confirms that the conditional ATEs are essentially not interpretable.

Figure 7: Type-conditional treatment effects

(a) $ATE(k, g)$



(b) $ATE(k)$ and $ATT(k)$



Notes: The bars show the type-conditional average treatment effects (ATEs), conditional on both first- and second-stage types in panel (a), and after aggregating over the second-stage types in panel (b). The colors in both panels correspond to ATEs in different years: 2013 (red), 2014 (green), 2015 (blue). The error bars are 90% confidence intervals obtained by bootstrap. In panel (b) the filled bar shows the ATE, while the black outline shows the average treatment on the treated (ATT, obtained by aggregating only over those who received training) for the same year.

Panel (b) of Figure 7 displays aggregate treatment effects conditional on first-stage types only, $ATE(k)$, obtained by summing over g conditional $ATE(k,g)$ weighted by $\pi(g|k)$. The empty black-outlined bars are the $ATT(k)$, obtained in the same way, but using weights $\pi(g|k,d=1)$. Recall, the first-stage classification orders workers by increasing abilities, a source of heterogeneity that determines wages independently of adult training. Generally the $ATE(k)$'s are small — around one percent or less — though for $k=2$ they are closer to three percent in 2014 and 2015, but imprecisely estimated. The $ATT(k)$'s are only marginally greater than the $ATE(k)$'s. This suggests workers are not selecting into training based on their *ex post* wage returns. Generally, the picture in Figure 7 suggests small positive wage returns to training for most individuals of less than 1%, with a small number (around 12%) of individuals enjoying higher wage returns of about 3%. Lastly, note that the 2013 returns to training vanish after aggregating within the first-stage types (k).

Finally, we aggregate across both types k and g to obtain a variety of treatment effects summarizing the whole sample, which are presented in Table 6. The top rows show the results obtained when the outcome is log-wage in levels. The bottom rows show results when the outcome is the difference in log-wages between pre- (2013) and post-treatment (2014 and 2015) periods.

The first three columns are estimates of the average treatment effects weighted differently, ATE, ATT and LATE (equations (4), (5) and (7)). These are plug-in estimates using our estimates of the structural parameters (treatment probabilities and density means and variances).

Then, in subsection 2.2 we calculated two asymptotic biases on OLS and 2SLS estimators, for a well specified model satisfying Assumptions 1-5: B_{OLS} and B_{IV} (equations (6) and (8)). We also show the plug-in estimates in two different columns.

Let then

$$b_{OLS} = ATT + B_{OLS}, \quad b_{IV} = LATE + B_{IV},$$

denote the corresponding plug-in estimates of the OLS and IV parameters

$$\frac{\mathrm{Cov}(y,d)}{\mathrm{Var}(d)} \quad \text{and} \quad \frac{\mathrm{Cov}(y,z)}{\mathrm{Cov}(d,z)}.$$

Finally, let $\widehat{b}_{OLS}$ and $\widehat{b}_{IV}$ denote the standard OLS and the IV estimates obtained by replacing the preceding population variances by sample variances. While for OLS the plug-in and analog estimates coincide ($\widehat{b}_{OLS} = b_{OLS}$), for IV there is an additional bias arising because, in the sample, pre-treatment wages are correlated with treatment and instrument, and post-treatment wages are correlated with the instrument given treatment, albeit minutely. The plug-in estimator of the IV population parameter imposes the model's assumptions. The 2SLS estimator does not. The OLS and IV estimates and their

Table 6: Aggregate treatment effects

	ATE	ATT	LATE	$\widehat{b}_{OLS} = b_{OLS}$	B_{OLS}	$\widehat{b}_{IV}$	b_{IV}	B_{IV}	$\widehat{b}_{IV} - b_{IV}$
Log-wage levels									
2013	0.003	0.002	0.006	0.158	0.156	0.179	0.188	0.182	-0.011
	(0.004)	(0.004)	(0.005)	(0.027)	(0.029)	(0.063)	(0.060)	(0.061)	(0.021)
2014	0.009	0.010	0.012	0.164	0.153	0.219	0.199	0.187	0.022
	(0.006)	(0.008)	(0.008)	(0.027)	(0.035)	(0.063)	(0.060)	(0.061)	(0.023)
2015	0.009	0.011	0.010	0.167	0.157	0.216	0.204	0.194	0.010
	(0.007)	(0.006)	(0.009)	(0.027)	(0.029)	(0.063)	(0.060)	(0.061)	(0.024)
Log-wage changes									
'14 vs '13	0.009	0.010	0.012	0.008	-0.002	0.040	0.015	0.003	0.025
	(0.006)	(0.008)	(0.008)	(0.009)	(0.022)	(0.017)	(0.011)	(0.013)	(0.023)
'15 vs '13	0.009	0.011	0.010	0.012	0.001	0.037	0.020	0.010	0.017
	(0.007)	(0.006)	(0.009)	(0.008)	(0.024)	(0.021)	(0.011)	(0.009)	(0.024)

Notes: (1) The ATE, ATT and LATE estimates for 2013 log-wage levels are zero by Assumption 5. To obtain the (nonzero) values in the top rows of the table, we compute $\mu_1(h, d)$ as mean log-wages weighted by posterior probabilities separately for trained and untrained workers. For log-wage differences, the ATE, ATT and LATE refer to $\mu_t(h, d) - \mu_1(h)$. (2) Standard errors are in parentheses, calculated as the standard deviation of the parameter estimates from 500 weighted-likelihood bootstrap repetitions. (3) $\hat{b}_{OLS}$ and $\hat{b}_{IV}$ are "naive" estimates obtained using ordinary least squares (OLS) and two-stage least squares (IV). b_{OLS} and b_{IV} are our model analogues of these estimates, calculated using the formulas in subsection 2.2.

associated biases, are displayed in Table 6.

We find similar-sized estimates of ATE, ATT, and LATE, of around 1%, with a big bootstrap standard error. The treatment effects calculated for wages in 2013 are much lower than those calculated for wages in 2014 and 2015, which is consistent with our identifying restriction.

The biases resulting from heterogeneous treatment and counterfactual wage levels, B_{OLS} and B_{IV}, are of the same order of magnitude as the respective OLS and IV estimates, $\hat{b}_{OLS}$ and $\hat{b}_{IV}$. This was expected: we already know that there is a lot of heterogeneity in wage trajectories.

The specification errors on the IV estimator ($\widehat{b}_{IV} - b_{IV}$) are small in comparison. They are similar in magnitude to treatment effects, but with a much larger bootstrap standard error. Therefore, we do not reject the model's assumptions.

In the bottom part of Table 6 we show the difference-in-difference (DiD) decomposition. Under the identifying restrictions that wages do not depend on the instrument given treatment and that pre-treatment wages do not depend on the treatment, DiD treatment effects and level treatment effects are identical; but the decompositions differ. We see that the bias due to heterogeneous trends — i.e. due to violation of the common trend assumption — is negligible for OLS (B_{OLS}) and is slightly bigger for IV (B_{IV}). It is therefore likely that the sizable IV estimate for wage differences (4%) is for a large part

(maybe half of it) just noise.

All this evidence suggests that the effect of adult training on wages is very small, quasi undetectable.

5 Conclusion

In this article, we developed and demonstrated the empirical use of a novel methodology for estimating treatment effects that allows for unobserved heterogeneity. The identification of conditional treatment effects given latent types (ATE, ATT, and LATE) is rendered possible by a combination of nonparametric difference-in-difference and instrumental-variable inference. Conventional monotonicity or common trend assumptions are not required for identification. In addition, we allow outcome variables (wages) to be Markovian given treatment and latent type. By assuming discrete types, we permit unobserved heterogeneity to condition observed outcomes, treatments, and instruments in a very general way. For example, no form of linearity nor homoscedasticity is required in contrast with factor models. This also allows us to base the estimation of a flexible parametric form of the model on the EM algorithm. Our method is generally applicable to other policy evaluation problems. In our application using novel French data on training and wages, we find that formal training has a small positive effect on wages, around 1% on average, except for a small fraction of workers for whom we find treatment effects of around 3%.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Abbring, J. H. and G. J. v. d. Berg (2003): "The Nonparametric Identification of Treatment Effects in Duration Models," *Econometrica*, 71, 1491–1517.

Acemoglu, D. and J.-S. Pischke (1998): "Why Do Firms Train? Theory and Evidence," *Quarterly Journal of Economics*, 113, 79–119.

——— (1999): "The Structure of Wages and Investment in General Training," *Journal of Political Economy*, 107, 539–572.

Allman, E. S., C. Matias, and J. A. Rhodes (2009): "Identifiability of parameters in latent structure models with many observed variables," *Annals of Statistics*, 37, 3099–3132.

Ashenfelter, O. C. (1978): "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47–57.

Attanasio, O., A. Kugler, and C. Meghir (2011): "Subsidizing Vocational Training for Disadvantaged Youth in Colombia: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, 3, 188–220.

Ba, B. A., J. C. Ham, R. J. LaLonde, and X. Li (2017): "Estimating (Easily Interpreted) Dynamic Training Effects from Experimental Data," *Journal of Labor Economics*, 35, 149–200.

Ballot, G., F. Fakhfakh, and E. Taymaz (2006): "Who Benefits from Training and R&D, the Firm or the Workers?" *British Journal of Industrial Relations*, 44, 473–495.

Bartel, A. P. (1995): "Training, Wage Growth, and Job Performance: Evidence from a Company Database," *Journal of Labor Economics*, 13, 401–425, publisher: [University of Chicago Press, Society of Labor Economists, NORC at the University of Chicago].

Biernacki, C., G. Celeux, and G. Govaert (2000): "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Blundell, R., L. Dearden, C. Meghir, and B. Sianesi (1999): "Human capital investment: the returns from education and training to the individual, the firm and the economy," *Fiscal Studies*, 20, 1–23.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016a): "Estimating multivariate latent-structure models," *Annals of Statistics*, 44, 540–563.

——— (2016b): "Non-parametric estimation of finite mixtures from repeated measurements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 211–229.

——— (2017a): "Nonparametric estimation of non-exchangeable latent-variable models," *Journal of Econometrics*, 201, 237–248.

Bonhomme, S., T. Lamadon, and E. Manresa (2017b): "A Distributional Framework for Matched Employer Employee Data," mimeo, University of Chicago.

——— (2019): "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 87, 699–739.

Bonhomme, S. and U. Sauder (2011): "Recovering Distributions in Difference-in-Differences Models: A Comparison of Selective and Comprehensive Schooling," *Review of Economics and Statistics*, 93, 479–494.

Bonnal, L., D. Fougere, and A. Serandon (1997): "Evaluating the Impact of French Employment Policies on Individual Labour Market Histories," *Review of Economic Studies*, 64, 683–713.

Booth, A. L. (1993): "Private Sector Training and Graduate Earnings," *Review of Economics and Statistics*, 75, 164–170.

Brodaty, T., B. Crepon, and D. Fougere (2001): "Using Matching Estimators to Evaluate Alternative Youth Employment Programs : Evidence from France, 1986-1988," in *Econometric Evaluations of Labour Market Policies*, ed. by M. Lechner and F. Pfeiffer, Physica, Heidelberg, 2000-25, 85–124.

Caliendo, M., D. A. Cobb-Clark, H. Seitz, and A. Uhlendorff (2016): "Locus of Control and Investment in Training," SOEPpapers on Multidisciplinary Panel Data Research 890, DIW Berlin, The German Socio-Economic Panel (SOEP).

Callaway, B. and T. Li (2019): "Quantile treatment effects in difference in differences models with panel data," *Quantitative Economics*, 10, 1579–1618.

Card, D., J. Kluve, and A. Weber (2010): "Active Labour Market Policy Evaluations: A Meta-Analysis," *Economic Journal*, 120, 452–477.

——— (2018): "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations," *Journal of the European Economic Association*, 16, 894–931.

Cardoso, J. F. (1989): "Sources separation using higher order moments," *Proc. Internat. Conf. Acoust. Speech Signal Process.*, 2109–2112.

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–2781.

Carneiro, P. and S. Lee (2009): "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 149, 191–208.

Crepon, B., M. Ferracci, G. Jolivet, and G. J. van den Berg (2009): "Active Labor Market Policy Effects in a Dynamic Setting," *Journal of the European Economic Association*, 7, 595–605.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931.

de Chaisemartin, C. and X. d'Haultfoeuille (2020): "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110, 2964–2996.

Dearden, L., H. Reed, and J. V. Reenen (2006): "The Impact of Training on Productivity and Wages: Evidence from British Panel Data," *Oxford Bulletin of Economics and Statistics*, 68, 397–421.

Fialho, P., G. Quintini, and M. Vandeweyer (2019): "Returns to different forms of job related training," Tech. Rep. 231, OECD Social, Employment and Migration Working Papers.

Fraley, C. and A. E. Raftery (1998): "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis," *The Computer Journal*, 41, 578–588.

Freyaldenhoven, S., C. Hansen, and J. M. Shapiro (2019): "Pre-event Trends in the Panel Event-Study Design," *American Economic Review*, 109, 3307–3338.

Gerfin, M. and M. Lechner (2002): "A Microeconometric Evaluation of the Active Labour Market Policy in Switzerland," *Economic Journal*, 112, 854–893.

Goux, D. and E. Maurin (2000): "Returns to firm-provided training: evidence from French worker-firm matched data," *Labour Economics*, 7, 1–19.

Grip, A. D. and J. Sauermann (2012): "The Effects of Training on Own and Co-worker Productivity: Evidence from a Field Experiment," *Economic Journal*, 122, 376–399.

Gritz, R. M. (1993): "The impact of training on the frequency and duration of employment," *Journal of Econometrics*, 57, 21–51.

Haelermans, C. and L. Borghans (2012): "Wage Effects of On-the-Job Training: A Meta-Analysis," *British Journal of Industrial Relations*, 50, 502–528.

Heckman, J. J., R. J. Lalonde, and J. A. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 3, 1865–2097.

Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

Henry, M., Y. Kitamura, and B. Salanie (2014): "Partial identification of finite mixtures in econometric models," *Quantitative Economics*, 5, 123–144.

Hu, Y. (2008): "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 144, 27–61.

——— (2015): "Microeconomic models with latent variables: Applications of measurement error models in empirical industrial organization and labor economics," Tech. rep., Cemmap, Working Papers, CWP03/15.

——— (2017): "The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics," *Journal of Econometrics*, 200, 154–168.

Hu, Y. and S. M. Schennach (2008): "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, 76, 195–216.

Hu, Y. H. and M. Shum (2012): "Nonparametric identification of dynamic models with unobserved state variables," *Journal of Econometrics*, 171, 32–44.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Kasahara, H. and K. Shimotsu (2009a): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77, 135–175.

——— (2009b): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77, 135–175.

Kluve, J., H. Schneider, A. Uhlendorff, and Z. Zhao (2012): "Evaluating continuous training programmes by using the generalized propensity score," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 175, 587–617.

Konings, J. and S. Vanormelingen (2015): "The Impact of Training on Productivity and Wages: Firm-Level Evidence," *Review of Economics and Statistics*, 97, 485–497.

Krueger, A. and C. Rouse (1998): "The Effect of Workplace Education on Earnings, Turnover, and Job Performance," *Journal of Labor Economics*, 16, 61–94.

LaLonde, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Leuven, E. and H. Oosterbeek (2008): "An alternative approach to estimate the wage returns to private-sector training," *Journal of Applied Econometrics*, 23, 423–434.

Li, F. and F. Li (2019): "Double-Robust Estimation in Difference-in-Differences with an Application to Traffic Safety Evaluation," *arXiv:1901.02152 [stat]*, arXiv: 1901.02152.

Lynch, L. M. (1992): "Private-Sector Training and the Earnings of Young Workers," *American Economic Review*, 82, 299–312.

McCall, B., J. Smith, and C. Wunsch (2016): "Government-Sponsored Vocational Education for Adults," in *Handbook of the Economics of Education*, Elsevier, vol. 5, 479–652.

Newton, M. A. and A. E. Raftery (1994): "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society: Series B (Methodological)*, 56, 3–26.

O'Hagan, A., T. B. Murphy, L. Scrucca, and I. C. Gormley (2019): "Investigation of parameter uncertainty in clustering using a Gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap," *Computational Statistics*, 34, 1779–1813.

Parent, D. (1999): "Wages and Mobility: The Impact of Employer-Provided Training," *Journal of Labor Economics*, 17, 298–317.

Pischke, J.-S. (2001): "Continuous training in Germany," *Journal of Population Economics*, 14, 523–548.

Ridder, G. (1986): "An Event History Approach to the Evaluation of Training, Recruitment and Employment Programmes," *Journal of Applied Econometrics*, 1, 109–126.

Rodriguez, J., F. Saltiel, and S. S. Urzua (2018): "Dynamic Treatment Effects of Job Training," NBER Working Papers 25408, National Bureau of Economic Research.

Sandvik, J., R. Saouma, N. Seegert, and C. T. Stanton (2021): "Treatment and Selection Effects of Formal Workplace Mentorship Programs," Working Paper 29148, National Bureau of Economic Research.

Sant'Anna, P. H. C. and J. Zhao (2020): "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, 219, 101–122.

Sasaki, Y. (2015): "Heterogeneity and selection in dynamic panel data," *Journal of Econometrics*, 188, 236–249.

Schoene, P. (2004): "Why is the Return to Training So High?" *Labour*, 18, 363–378.

Shiu, J.-L. and Y. Hu (2013): "Identification and estimation of nonlinear dynamic panel data models with unobserved covariates," *Journal of Econometrics*, 175, 116–131.

A Proof of the Identification Theorem

The identification proof has four steps.

Step 1: Identifying restrictions. Consider first the joint probability $p(z, d, w_1, w_2, w_3)$ of treatment $d_i = d$, instrument $z_i = z$, and wages $w_{i1} \leq w_1$ (before treatment) and $w_{i2} = w_2, w_{i3} \leq w_3$ (after treatment). We now drop index i to lighten notation. Mixing over unobserved types, for any wage $w_2 \in \mathcal{W}_2(d)$ — such that $f_2(w_2|h, d) \neq 0$ at least for one h — we can write

$$p(z, d, w_1, w_2, w_3) = \sum_{h: f_2(w_2|h,d)\neq 0} \pi(h, z, d)\, f_2(w_2|h, d)\, F_{1|2}(w_1|w_2, h, d)\, F_{3|2}(w_3|w_2, h, d),$$

where $F_{1|2}$ and $F_{3|2}$ denote distribution functions and f_2 a density. Notice how we first condition on w_{i2}. The sum is therefore over the values of h such that $f_2(w_2|h, d) \neq 0$.

Let us consider a grid of N wages w_1 and M wages w_3, including maximal wages $\overline{w}_1, \overline{w}_3$. Then, for any value of (z, d, w_2), we can store these probabilities $p(.)$ in a matrix

$$P(z, d, w_2) = \left[p(z, d, w_1, w_2, w_3)\right]_{w_1 \times w_3},$$

where the subscript $w_1 \times w_3$ means that the values of w_1 index rows and those of w_3 index columns. Let

$$D(z, d, w_2) = \operatorname{diag}\left[\pi(h, z, d)\, f_2(w_2|h, d)\right]_{h: f_2(w_2|h,d)\neq 0}$$

be the diagonal matrix with $\pi(h, z, d)\, f_2(w_2|h, d)$ in the hth diagonal entry, keeping only the values of h such that $f(w_2|h, d) \neq 0$. Let also $G_1(d, w_2) = \left[F_{1|2}(w_1|w_2, h, d)\right]_{w_1 \times h}$ denote the matrix of pre-treatment wage probabilities, with w_1 indexing rows and h indexing columns. Similarly, let $G_2(d, w_2) = \left[F_{3|2}(w_3|w_2, h, d)\right]_{w_3 \times h}$ be the post-treatment matrix. Again, the values of h indexing columns are only those such that $f_2(w_2|h, d) \neq 0$. Note that the first row of G_1, G_2 is a row of ones. Finally, In matrix notation, we then have, for every (w_2, z, d),

$$P(z, d, w_2) = G_1(d, w_2)\, D(z, d, w_2)\, G_2(d, w_2)^\top.$$

The number of columns of $G_1(d, w_2)$ and $G_2(d, w_2)$ and the dimensions of $D(z, d, w_2)$ vary with w_2, as we keep only those values of h such that $f_2(w_2|h, d) \neq 0$ in their construction. But, we do not know what they are *a priori*.

Step 2: Identification given treatment d and first post-treatment wage w_2. We first fix a value d of the treatment variable and a wage $w_2 \in \mathcal{W}_2(d)$. The previous step shows that, for all d, w_2, there are two observable matrices, $P(0, d, w_2)$ and $P(1, d, w_2)$, with the same algebraic structure. Importantly, $G_1(d, w_2)$ and $G_2(d, w_2)$ are independent

of z as wages are independent of the instrument given treatment and type (Assumption 1). Under Assumption 3, $G_1(d, w_2)$ and $G_2(d, w_2)$ are full-column rank, and under Assumption 2 the matrix $D(0, d, w_2)$ is invertible for all $w_2 \in \mathcal{W}_2(d)$. Also, by Assumption 4 all diagonal entries of $D(1, d, w_2)D(0, d, w_2)^{-1}$ are distinct. Finally, all first row entries of $G_1(d, w_2)$ and $G_2(d, w_2)$ contain ones. We deduce from the following lemma the identification of $G_1(d, w_2)$, $D(z, d, w_2)$ and $G_2(d, w_2)$.

Lemma 2 (Standardization)**.** *Let $P(0), P(1) \in \mathbb{R}^{N\times M}$ be two matrices with similar algebraic structure: $P(z) = G_1 D(z) G_2^\top$, $z \in \{0, 1\}$, where $G_1, G_2, D(z)$ satisfy the following restrictions: i) $G_1 \in \mathbb{R}^{N\times H}$ and $G_2 \in \mathbb{R}^{M\times H}$ are two full column-rank; ii) $D(z) \in \mathbb{R}^{H\times H}$ are diagonal; iii) $D(0)$ is non singular; iv) all diagonal entries of $D(1)D(0)^{-1}$ are distinct; v) the first rows of G_1 and G_2 are made of ones. Then, $G_1, G_2, D(0)$ and $D(1)$ are uniquely determined by $P(0), P(1)$.*

Proof. Matrix $P(0)$ has rank H and there exists a singular value decomposition: $P(0) = U\Lambda V^\top$, where $U \in \mathbb{R}^{N\times N}$ and $V \in \mathbb{R}^{M\times M}$ are nonsingular orthogonal matrices with $U^\top U = I_N$, $V^\top V = I_M$ and $\Lambda \in \mathbb{R}^{N\times M}$ is a rectangular diagonal matrix with non-negative real numbers on the diagonal. The number of non-zero diagonal entries in Λ is equal to H. Let $\Lambda_1 \in \mathbb{R}^{H\times H}$ be the square diagonal matrix containing the non-zero singular values, and let $U = (U_1, U_2)$ and $V = (V_1, V_2)$ partition the columns of Λ accordingly, so that $P(0) = U_1\Lambda_1 V_1^\top$.

Next, using the singular value decomposition of $P(0)$, we have

$$\Lambda_1^{-1} U_1^\top P(0) V_1 = \Lambda_1^{-1} U_1^\top U_1 \Lambda_1 V_1^\top V_1 = I_H.$$

Hence, $\Lambda_1^{-1} U_1^\top G_1 D(0) G_2^\top V_1 = I_H$. Define $W = \Lambda_1^{-1} U_1^\top G_1 \in \mathbb{R}^{H\times H}$. The matrix W is thus non singular and $W^{-1} = D(0) G_2^\top V_1$.

Now, we also find that

$$\Lambda_1^{-1} U_1^\top P(1) V_1 = \Lambda_1^{-1} U_1^\top G_1 D(1) G_2^\top V_1 = W D(1) D(0)^{-1} W^{-1}.$$

The diagonal entries of $D(1)D(0)^{-1}$ being distinct, they are uniquely determined as the eigenvalues of the matrix $\Lambda_1^{-1} U_1^\top P(1) V_1$. However, eigenvectors are determined only up to a multiplicative constant. So, let $\widehat{W}$ be one matrix of eigenvectors. There exists a non-singular diagonal matrix Δ such that $\widehat{W} = W\Delta = \Lambda_1^{-1} U_1^\top G_1 \Delta$. Then, $\Lambda_1 \widehat{W} = U_1^\top G_1 \Delta$.

It is not true that $U_1 U_1^\top = I_N$ because the columns of U_1 are orthogonal but not its rows. However, since the columns of U are orthogonal vectors,

$$U_2^\top P(0) = U_2^\top U_1 \Lambda_1 V_1^\top = 0_{(N-H)\times M}.$$

Hence, $U_2^\top G_1 D(0) G_2^\top = 0_{(N-H)\times M}$. As $D(0) G_2^\top \in \mathbb{R}^{H\times M}$ is a full row-rank, it follows that

$U_2^\top G_1 = 0_{(N-H)\times H}$. A similar argument implies that $P(0)V_2 = 0$ since $V_1^\top V_2 = 0$. Now, since $G_1 D(0)$ has rank H, it follows that $G_2^\top V_2 = 0_{H\times(M-H)}$. From $U_2^\top G_1 \Delta = 0_{(N-H)\times H}$, we deduce that

$$\begin{pmatrix} \Lambda_1 \widehat{W} \\ 0_{(N-H)\times H} \end{pmatrix} = U^\top G_1 \Delta.$$

Hence,

$$U_1 \Lambda_1 \widehat{W} = (U_1, U_2) \begin{pmatrix} \Lambda_1 \widehat{W} \\ 0_{(N-H)\times H} \end{pmatrix} = U U^\top G_1 \Delta = G_1 \Delta.$$

Since G_1 contains a row of ones, then the last equality implies that the diagonal of Δ is identified by the first row of $U_1 \Lambda_1 \widehat{W}$. Then $G_1 = U_1 \Lambda_1 \widehat{W} \Delta^{-1}$ follows.

Lastly, we have $\Delta \widehat{W}^{-1} = W^{-1} = D(0) G_2^\top V_1$. Applying the same argument as above, we have that

$$\begin{aligned} W^{-1} V_1^\top &= \left(D(0) G_2^\top V_1, 0_{H\times(M-H)} \right) \begin{pmatrix} V_1^\top \\ V_2^\top \end{pmatrix} \\ &= \left(D(0) G_2^\top V_1,\ D(0) G_2^\top V_2 \right) V^\top \\ &= D(0) G_2^\top V V^\top \\ &= D(0) G_2^\top. \end{aligned}$$

In the same way as above, the first row of G_2 is made of ones, it follows that $D(0)$ and G_2 are identified. Hence $D(1)$ is also identified. □

Step 3: Common labelling given d. In the previous step, we have identified

$$D(1, d, w_2) D(0, d, w_2)^{-1} = \operatorname{diag} \left[\frac{\pi(h, 1, d)}{\pi(h, 0, d)} \right]_{h: f_2(w_2|h,d)\neq 0}.$$

By Assumption 4, these eigenvalues are all different (and independent of w_2). One can thus relabel groups for each d so that the labelling is consistent for all possible choices of w_2. This also allows to identify the different supports $\mathcal{W}_2(h, d)$.

Step 2 can be done for all wages w_2 in the joint support $\mathcal{W}_2(d) = \bigcup_h \mathcal{W}_2(h, d)$. Thus, we can sum $D(0, d, w_2)$ and $D(1, d, w_2)$ over w_2 and eliminate $f_2(w_2|h, d)$ (which sums to one on its support). This identifies $\pi(0, h, d)$ and $\pi(1, h, d)$ for all h. Knowing $\pi(h, z, d)$ and $D(z, d, w_2)$, we identify $f_2(w_2|h, d)$.

Since $F_{1|2}(w_1|w_2, h, d)$ is already identified, then the Law of Total Probability implies that $F_1(w_1|h, d)$ is identified. Also, we can take the grid of wages w_1 as fine as we want. Bayes' formula therefore implies that $F_{2|1}(w_2|w_1, h, d)$ is also identified.

Step 4: Common labelling across treatments. It remains to align the groupings across treatments. This is done by remarking that $F_1(w_1|h)$ is independent of d (Assump-

tion 5) and therefore, can be used to make sure that the same groups have identical labels across treatments.
Q.E.D.

B Sequential EM-algorithm formulas

B.1 Stage 1

In stage 1 wages (and types) are independent of training, d, and information z. We denote first-stage types by k, and the number of types is K. We assume that log-wages are normal and denote log-wage in period t by w_t. The difference with the model described above is that μ and σ depend only on the first-stage type k, and do not depend on d. To avoid ambiguity, we use an upper bar to distinguish the first-stage from the second-stage variables and parameters,

$$\begin{aligned} w_1 &= \bar{\mu}_1(k) + u_1, \quad u_1 \sim N\left(0, \bar{\sigma}_1^2(k)\right), \\ w_t &= \bar{\mu}_t(k) + u_t, \quad u_t \sim N\left(\bar{\rho} u_{t-1}, \bar{\sigma}_t^2(k)\right), \quad t = 2, 3. \end{aligned}$$

E-step. The complete individual likelihood in stage 1 has a simplified form (depending only on k), that is,

$$\bar{\ell}_{ik}(\beta) = \bar{\pi}(k)\, \bar{f}_1(w_{i1}|k)\, \bar{f}_{2|1}(w_{i2}|w_{i1}, k)\, \bar{f}_{3|2}(w_{i3}|w_{i2}, k) \tag{11}$$

The posterior probability of worker i to be of type k given data (i.e., the conditional probability of k knowing i, also called *responsibility*), denoted $\bar{p}_{ik}$, can be computed with the help of contributions to likelihood, using Bayes' rule. Let $\beta^{(m)}$ denote an estimate of the parameters at the end of iteration m. More precisely, we have,

$$\bar{p}_{ik}^{(m)} \equiv \frac{\bar{\ell}_{ik}(\beta^{(m)})}{\sum_k \bar{\ell}_{ik}(\beta^{(m)})}. \tag{12}$$

M-step. We update the parameters sequentially as follows, using the following sequential procedure.

1. Update pre-treatment wage distribution parameters $\overline{\mu}, \overline{\sigma}^2$ given current iteration $\bar{\rho}^{(m-1)}$ of the AR parameter as

$$\bar{\mu}_t^{(m)}(k) = \frac{\sum_i \bar{p}_{ik}^{(m)} w_{it}}{\sum_i \bar{p}_{ik}^{(m)}},$$

and, with $\bar{u}_{itk}^{(m)} = w_{it} - \bar{\mu}_t^{(m)}(k)$ for $t = 1$,

$$(\bar{\sigma}_1^2)^{(m)}(k) = \frac{\sum_i \bar{p}_{ik}^{(m)} \left(\bar{u}_{i1k}^{(m)}\right)^2}{\sum_i \bar{p}_{ik}^{(m)}},$$

and for $t = 2, 3$,

$$(\bar{\sigma}_t^2)^{(m)}(k) = \frac{\sum_i \bar{p}_{ik}^{(m)} \left[\bar{u}_{itk}^{(m)}(k) - \bar{\rho}^{(m-1)} \bar{u}_{i,t-1,k}^{(m)}\right]^2}{\sum_i \bar{p}_{ik}^{(m-1)}}.$$

2. Then update $\bar{\rho}$ as follows,

$$\bar{\rho}^{(m)} = \frac{\sum_i \sum_k \bar{p}_{ik}^{(m)} \left(\frac{\bar{u}_{i1k}^{(m)} \bar{u}_{i2k}^{(m)}}{(\bar{\sigma}_2^2)^{(m)}(k)} + \frac{\bar{u}_{i2k}^{(m)} \bar{u}_{i3k}^{(m)}}{(\bar{\sigma}_3^2)^{(m)}(k)} \right)}{\sum_i \sum_k \bar{p}_{ik}^{(m)} \left(\frac{(\bar{u}_{i1k}^{(m)})^2}{(\bar{\sigma}_2^2)^{(m)}(k)} + \frac{(\bar{u}_{i2k}^{(m)})^2}{(\bar{\sigma}_3^2)^{(m)}(k)} \right)}.$$

   The standard EM procedure would have all $\overline{\mu}, \overline{\sigma}^2$ and $\overline{\rho}$ estimated by weighted nonlinear least squares. By simplifying the M-estimation in this way, we do not obtain efficient M-step updates but the sequential EM algorithm keeps increasing the likelihood at each iteration.

3. Finally, we update $\bar{\pi}$ as the mean posterior probability across all workers,

$$\bar{\pi}^{(m)}(k) = \frac{1}{N} \sum_i \bar{p}_{ik}^{(m)}.$$

We continue iterating between these steps until the algorithm converges. The key results from the first stage are the final posterior probabilities, $\bar{p}_{ik}$, which we use as weights throughout the second stage to "allocate" workers to types.

B.2 Stage 2

To understand our 2-stages procedure, it can be helpful to consider the hypothetical case of a perfect (or hard) classification in stage 1. If each individual i belonged to only one group (or type) with probability 1, then, we would run stage 2 on each type k, only including those individuals classified as that type. With our soft classification, the posteriors $\bar{p}_{ik}$ can take any value between zero and one. Therefore, each observation i can contribute to the estimation of the model of several types in stage 2.

We assume now that the distribution of log-wages in period t, still denoted w_t, now

depend on (k, g) and treatment d. Wages are given by the following expressions,

$$w_1 = \mu_1(k, g) + u_1, \quad u_1 \sim N\left(0, \sigma_1^2(k, g)\right) \tag{13}$$

$$w_t = \mu_t(k, g, d) + u_t, \quad u_t \sim N\left(\rho u_{t-1}, \sigma_t^2(k, g, d)\right), \quad t = 2, 3 \tag{14}$$

The complete individual likelihood for stage 2 is now given by:

$$\ell_{ikg}(\beta) = \pi(g, k, z_i, d_i)\, f_1(w_{i1}|k, g)\, f_{2|1}(w_{i2}|w_{i1}, k, g, d_i)\, f_{3|2}(w_{i3}|w_{i2}, k, g, d_i) \tag{15}$$

In stage 2, we run the following procedure for each type $k \in K$ obtained in stage 1. As in stage 1, we iterate between an E-step (in which we update the posterior probabilities) and an M-step (in which we maximize the likelihood given the posteriors from the E-step). But we use the expression of p_{ik}, obtained in the first stage, to compute the posterior probabilities of all (k, g)s.

E-step. In the E-step, at the m-th iteration, we update the posterior probabilities as follows,

$$p_{ig|k}^{(m)} = \frac{\ell_{ikg}(\beta^{(m)})}{\sum_g \ell_{ikg}(\beta^{(m)})}. \tag{16}$$

Let also $p_{ikg}^{(m)} = \bar{p}_i(k) p_{ig|k}^{(m)}$ (using the estimated posterior probabilities $\bar{p}_i(k)$ from the first stage).

M-step. In the M-step we update the parameters of the likelihood function sequentially.

1. For $t = 1$:

$$\mu_1^{(m)}(k, g) = \frac{\sum_i p_{ikg}^{(m)} w_{i1}}{\sum_i p_{ikg}^{(m)}}, \tag{17}$$

$$(\sigma_1^2)^{(m)}(k, g) = \frac{\sum_i p_{ikg}^{(m)} (u_{i1kg}^{(m)})^2}{\sum_i p_{ikg}^{(m)}}, \tag{18}$$

with $u_{i1kg}^{(m)} = w_{i1} - \mu_1^{(m)}(k, g)$.

2. Then, for $t = 2, 3$,

$$\mu_t^{(m)}(k,g,d) = \frac{\sum_{\{i:d_i=d\}} p_{ikg}^{(m)} \left[w_{it} - \rho^{(m-1)} u_{i,t-1,kgd}^{(m)} \right]}{\sum_{\{i:d_i=d\}} p_{ikg}^{(m)}}$$

$$(\sigma_t^2)^{(m)}(k,g,d) = \frac{\sum_{\{i:d_i=d\}} p_{ikg}^{(m)} \left[u_{itkgd}^{(m)} - \rho^{(m-1)} u_{i,t-1,kgd}^{(m)} \right]^2}{\sum_{\{i:d_i=d\}} p_{ikg}^{(m)}},$$

where $u_{itkgd}^{(m)} = w_{it} - \mu_t^{(m)}(k,g,d), t = 2, 3$.
Note that $\mu_t(k,g,d)$ now depends on ρ for $t = 2, 3$ because we impose $\mu_1(k,g,0) = \mu_1(k,g,1) = \mu_1(k,g)$, i.e., treatment d has no effect on pre-treatment wages, conditional on type (k,g). If we relaxed this constraint, the estimator $\mu_t(k,g,d)$ would always be a simple weighted average of w_{it}.

3. Denote $I(d) = \{i : d_i = d\}$, then, we can update the autoregressive parameter ρ as follows,

$$\rho^{(m)} = \frac{\sum_{k,g} \sum_{d\in\{0,1\}} \sum_{i\in I(d)} p_{ikg}^{(m)} \left(\frac{u_{i1kg}^{(m)} u_{i2kgd}^{(m)}}{(\sigma_2^2)^{(m)}(k,g,d)} + \frac{u_{i2kgd}^{(m)} u_{i3kgd}^{(m)}}{(\sigma_3^2)^{(m)}(k,g,d)} \right)}{\sum_{k,g} \sum_{d\in\{0,1\}} \sum_{i\in I(d)} p_{ikg}^{(m)} \left(\frac{(u_{i1kg}^{(m)})^2}{(\sigma_2^2)^{(m)}(k,g,d)} + \frac{(u_{i2kgd}^{(m)})^2}{(\sigma_3^2)^{(m)}(k,g,d)} \right)}.$$

4. Finally, the type-state probabilities $\pi(k,g,z,d)$ are estimated as the average of posterior probabilities

$$\pi^{(m)}(k,g,z,d) = \frac{1}{N} \sum_{\{i:z_i=z,d_i=d\}} p_{ikg}^{(m)}.$$