A Nonparametric Finite-Mixture Approach to Instrumented Difference-in-Differences, with an Application to Job Training

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

June 13, 2024

Abstract

We develop a finite-mixture framework for nonparametric difference-in-differences analysis with unobserved heterogeneity correlating treatment and outcome. Our framework includes an instrumental variable for the treatment, and we prove nonparametric identification. We can thus relax the single index and stationarity assumptions of Athey and Imbens (2006) at the cost of adding slightly more structure on unobserved heterogeneity. We apply our framework to evaluate the effect of on-the-job training on wages, using novel French linked employee-employer data. Estimating a parametric version of our model with the help of an EM-algorithm, we find small ATEs and ATTs on hourly wages, around 4% in the year of training, falling to under 2% in the following year.

Keywords: Difference-in-differences; Finite Mixtures; Unobserved Heterogeneity; Instrumental Variables; EM Algorithm; Wage Distributions; Training
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 with the average outcome of a control group before and after the implementation of the program of interest. For this estimator to identify the causal effect of the program, one must assume that, in the absence of the treatment, the outcomes of both groups would have followed parallel trends over time, at least conditional on observable pretreatment covariates (Sant'Anna and Zhao, 2020). Unfortunately, it is not always possible to observe all possible confounders to apply this framework plausibly. Researchers therefore often opt for an instrumental variable approach. An instrument is any form of random assignment to the treatment, and the IV estimator is then more or less informative depending on the degree and type of compliance (see Athey and Imbens, 2017, for a review).

In this paper, we propose a novel approach to be used with unobserved heterogeneity and heterogeneous treatment effects. We exploit recent advances in the identification of latent variables models (see Hu, 2017, for a survey) to develop a framework where the whole distribution of heterogeneous treatment effects is identified. To this end, we assume that unobserved heterogeneity can be well approximated by a discrete distribution. Discrete mixing is not essential (Hu and Shum, 2012), but simplifies both identification and estimation. We show that the components of the finite mixture are identified if there exists a variable that affects the treatment probability without directly affecting the outcome conditional on unobserved individual types. This variable can be assigned by a lottery (as in the classical intention-to-treat framework) or by an agent (a policymaker, or a firm in our application). In many cases, the treatment assignment is not randomized and may be correlated with individual characteristics that are unobserved by the econometrician. Although such setups are quite common, they have received little attention in the econometric literature. Our method aims to fill this gap.

Once the distribution of the treatment effect conditional on unobservable types has been identified, the conventional monotonicity assumption is no longer needed, nor is the common trend assumption. We can estimate treatment effects conditional on type, and then choose the weights as we want for aggregation across types. But there is no free lunch; the identification of our model requires satisfying completeness assumptions similar to those in the nonparametric IV literature (Newey and Powell, 2003; Newey, 2013). In particular, we need sufficient variation across unobserved types in the distribution of the outcome and the instrument, separately for the treated and the untreated.

Using 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. We use data on whether or not the worker has received information about training opportunities as

a pretreatment assignment ("intention to treat"). The availability of repeated observations of the outcome variable is crucial for identification, and to model outcome dynamics. As we observe wages more than twice, we allow for the outcome process to be autoregressive by modeling it as a Markovian process.

We estimate a flexible parametric specification using the Expectation-Maximization (EM) algorithm. Standard errors are obtained by bootstrap. The results show that treatment effects vary with type. All three ways of aggregating conditional ATEs (aggregate ATE, ATT, and LATE) yield similar estimates of 4% in the year of training, falling to under 2% in the year after training. We conclude that on-the-job training has a limited effect on wages, which does not last. The biases resulting from heterogeneous trends are found to be of a similar magnitude to the aggregate treatment effects. We also find that a sizable share of the bias on the IV estimator (around 1%-2%) reflects small-sample deviations from assumed restrictions in the population.

Our framework should be viewed as an alternative to the DiD framework (OLS or IV). We solve the nonparametric identification problem by assuming that the source of confoundedness lies in the presence of unobserved heterogeneity that we assume fixed over time. The DiD framework makes other assumptions. Athey and Imbens (2006) state them clearly: First, there should be a unique stochastic index determining all counterfactual outcomes. Second, the index should be a stationary stochastic process. The fixed unobserved heterogeneity assumption is a strong restriction. However, it is relatively straightforward to extend our framework to the hidden Markov case. But, more outcome observations would be required (see Hu and Shum, 2012), which is a problem in our application because increasing the number of periods also increases the risk of re-training.

The benefits of panel data in difference-in-differences 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 the 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 with a structural assumption on the way unobserved heterogeneity 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, and consider a more general finite mixture allowing the outcome variance to be conditional on unobserved heterogeneity.

[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 pretreatment 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.

The rest of the paper proceeds as follows. We first conclude the introduction with a discussion of the literature on training. Then, Section 2 presents the model. Section 3 discusses identification. Section 4 makes the links between our model's estimates and standard treatment effect parameters such as ATE, ATT, OLS and LATE. Section 5 describes our dataset and presents a preliminary, reduced-form econometric analysis using standard econometric methods. Section 6 presents and discusses the results obtained with our new framework. We finally conclude.

Literature on training. The literature on the effect of training (and active labor market programs) is vast. 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 or without controls respectively. 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 control 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; *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 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% with less restrictive choices.[3]

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 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 indeed find positive effects on both productivity and wages.[4]

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 impact 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 other empirical 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, 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 goal is to measure the causal impact of training on wages in periods $t = 2, 3$. Treatment d_i is a binary variable, training or not training, although the model and the identification proof encompass the case of treatment variables with any finite number of values. Specifically, we could allow for different types of training, for example by duration. We also single out, from all potential control or outcome variables, a variable $z_i \in \{0, 1\}$, which will play an important role in identification and that we call the instrument.

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

[4] 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: $h_i \in \{1, ..., H\}$. The distributions of all variables w_{it}, z_i and d_i potentially vary 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 condition wages and training. It is of course possible to first cluster the data, say by education, and run the study separately within each education group. 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_i and a set of available controls *ex post*.

We start by making the following basic restrictions on the structure of the model.

Assumption 1 (Model structure)**.** *The basic model structure satisfies the following restrictions:*

1. *The wage process $(w_{it})_{t\geq 1}$ is first-order Markov and bounded given the type, the instrument and the treatment.*

2. *The pretreatment wage is independent of the instrument and the treatment given the type: $w_{i1} \perp\!\!\!\perp z_i, d_i \mid h_i$.*

3. *The post-treatment wages are independent of the instrument given the type and the treatment: $(w_{it})_{t>1} \perp\!\!\!\perp z_i \mid h_i, d_i$.*

Allowing for residual wage autocorrelation (Condition 1) given unobserved heterogeneity is essential for empirical credibility. The workhorse model of the empirical wage dynamics literature assumes that the wage process is the sum of a fixed effect, a random walk and a transitory component, usually a stationary MA(0) or MA(1) process. If wages are i.i.d., then two wage observations are sufficient for identification. But we need three wage observations to identify the model for Markovian wages. They 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. We can easily allow for an autoregressive process of higher degree; but the Markov property is crucial for identification. The bounded support assumption is a simplification (see the proof sketch below).

Conditions 2 and 3 require that the special variable z_i should not determine wages once heterogeneity and training are controlled for. This exclusion restriction is fundamental for our identification argument. Of course, the instrument will be useful to the extent that it correlates with unobserved heterogeneity and/or treatment.

A useful way of thinking of the instrument is within a Roy model (see Heckman and Vytlacil, 2005; Carneiro and Lee, 2009; Carneiro et al., 2010, 2011). Let w_0, w_1 denote the potential outcomes (i.e., post-treatment wages) and let c be a random training cost. Then, training occurs ($d = 1$) if the expected return $\mathrm{E}[w_1 - w_0 \mid h]$ is greater than the

cost c. The instrument may enter this model in two different ways. Either z is a share of the cost, in addition to a fixed effect and an idiosyncratic component: $c = z + h + v$; or z is not a cost factor ($c = h + v$), but measures unobserved heterogeneity (z and h are correlated).

In our application, z_i will follow both interpretations. We will use the response to the questionnaire about whether the worker has received information about the availability of training sessions through any of the following channels: hierarchy, human resources, coworkers, or unions. We think of this variable as an intention to treat: receiving training information should encourage training. At the same time it is likely that some worker types will be given training information more often than others. Lastly, conditional on all worker heterogeneity, it seems reasonable to assume that training information has no causal effect on wages.

We assume that the pretreatment wage is independent of both the instrument and the treatment (Condition 2). 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 year before training, which makes predetermination more plausible.

Leaving the instrument aside, our model adopts a standard difference-in-differences structure. However, it is less restrictive in some important ways. The distributional change version in Athey and Imbens (2006) assumes that there exists unidimensional factors $(u_{it})_{t\geq 1}$ and monotone functions h_1 and h_{td}, for $t \geq 2, d \in \{0,1\}$, such that the pretreatment wage is $w_{i1} = h_1(u_{i1})$ and the counterfactual post-treatment wages are $w_{td} = h_{td}(u_{it})$. The indexes $(u_{it})_{t\geq 1}$ can be autocorrelated and correlated with the treatment d_i, possibly through unobserved heterogeneity. Thus, the model of Athey and Imbens is more restrictive in two ways. First, they assume a single index model for counterfactual wages. Second, the index process $(u_{it})_{t\geq 1}$ should be stationary given d_i (constant marginal distributions). We can relax the single index and the stationarity assumptions at the cost of adding more structure on unobserved heterogeneity. First, the dependence between wage innovations and treatment is mediated by latent types, which remain fixed over time. Second, we assume that there exists a special variable, which we call the instrument, and whose exact role in identification will be described in the next section.

3 Identification

The identification argument builds on the literature on the identification of latent variables models.[5]

Let $F_t(w_{it} \mid h, d)$ denote the distribution function for the marginal distribution of wages w_{it} given treatment and type. Let $F_{t|s}(w_{it} \mid w_{is}, h, d)$ denote the distribution function for the conditional distribution of w_{it} given w_{is} (we use $s = t \pm 1$). Corresponding densities are denoted with a lower-case f. These distributions can be discrete or continuous. Let $\mathcal{W}_2(h, d)$ be the support of $f_2(w_2 \mid h, d)$ (the set of wages w such that $f_2(w_2 \mid h, d) > 0$) 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 $d \in \{0, 1\}$.

The density of the individual data $(z_i, d_i, w_{i1}, w_{i2}, w_{i3})$ can be factored in the following way:

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

where, by Bayes' rule,

$$f_{1|2}(w_1 \mid w_2, h, d) = \frac{f_{2|1}(w_2 \mid w_1, h, d) f_1(w_1 \mid h)}{f_2(w_2 \mid h, d)},$$
$$f_2(w_2 \mid h, d) = \sum_h f_{2|1}(w_2 \mid w_1, h, d) f_1(w_1 \mid h).$$

The static case of i.i.d. wages given latent types and treatment can be seen as a particular case where the distribution of w_{i2} is a mass on w_1 ($\mathcal{W}_2(d) = \{w_1\}$).

We show that all the components of the right-hand side of equation (1) are identified under the following assumptions. These assumptions could be collected into one single completeness assumption, as they all play an essential role in proving that an operator relating observables to parameters is invertible.

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 of wages distributions)**.** *For all d, all $w_2 \in \mathcal{W}_2(d)$, and $t = 1, 3$, there exists grids of wages (w_t) such that the systems*

$$\Big\{F_{t|2}(w_t \mid w_2, h, d), \forall h : f_2(w_2 \mid h, d) \neq 0\Big\}$$

[5]See Cardoso, 1989; Bonhomme and Robin, 2009 for Independent Component Analysis, and for mixtures see Hall and Zhou (2003); Hu (2008); Allman et al. (2009); Kasahara and Shimotsu (2009); Hu and Shum (2012); Hu (2015); Bonhomme et al. (2016a,b); Hu (2017).

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 that is similar to the completeness assumptions in semiparametric IV models (see Theorem 2.4 of Newey and Powell, 2003, for instance). It generalizes the rank conditions for identification in Least Squares models.

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

Equivalently, $\pi(z = 1 \mid h, d)$ must vary across h for all d. If z is taking more than two values, then, separately for all d, we require the existence of a linear combination of $\frac{\pi(h,z,d)}{\pi(h,0,d)}$, for $z \neq 0$, satisfying the assumption.

Such an 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). This assumption is important to identify the type-specific wage distributions. Technically, these probability ratios will be identified as eigenvalues and the distribution components as eigenvectors. If two eigenvalues are the same, only linear combinations of the corresponding group distributions are identified.

Furthermore,

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

If the instrument is perfectly randomized — $\pi(h \mid z)$ is independent of z — then Assumption 4 means that different types must show different probabilities of complying. Or the instrument is not predictive of treatment, but it is not randomized and measures unobserved heterogeneity (different distributions of z in different groups).

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, $f_1(w_1 \mid h, d) = f_1(w_1 \mid h)$ and all densities $f_1(\cdot \mid h)$ are different.*

This assumption repeats parts of Condition 2 of Assumption 1 and adds that the $f_1(\cdot \mid h)$ are different for all h. It is used to recover a common labeling of groups across treatments as the identification of the model's distributions is first done separately for each value of the treatment.

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

The detailed proof of the identification theorem is in Appendix A. Here we sketch the proof with just two wage observations and no wage dynamics to emphasize the roles of the various assumptions. In the detailed proof, we consider the case of Markovian wages. The proof differs from that in Hu and Shum (2012), who require four wages for identification, but one can think of the instrument as a fourth measurement.

Sketch of the proof. The likelihood of the joint event $\{w_{i1} \leq w_1, w_{i2} \leq w_2, z_i = z, d_i = d\}$ is

$$p(z, d, w_1, w_2) = \sum_h \pi(h, z, d)\, F_1(w_1 \mid h)\, F_2(w_2 \mid h, d).$$

Thus, for each d, 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 is correlated with h but not with wages given h. Let us consider a grid of wages, and let us store the discretized function $F_1(w_1 \mid h)$ in a matrix $\mathbf{F}_1 = [F_1(w_1 \mid h)]_{w_1,h}$, where wage points index rows and latent types index columns. Similarly, let $\mathbf{F}_2(d) = [F_2(w_2 \mid h, d)]_{w_2,h}$. The first rows of $\mathbf{F}_1, \mathbf{F}_2(d)$ are made of ones if the first point of the grid is the maximal wage (bounded support). Next, let $P(z, d) = [p(z, d, w_1, w_2)]_{w_1,w_2}$ store the likelihood values in a matrix where w_1 indexes rows and w_2 indexes columns. Finally, let $D(z, d) = \text{diag}\,[\pi(1, z, d), ..., \pi(H, z, d)]$ be a diagonal matrix with the latent type probabilities $\pi(h, z, d)$ along the diagonal. Then, $P(z, d) = \mathbf{F}_1 D(z, d) \mathbf{F}_2(d)^\top$. Note that the number of types H is simply the rank of $P(z, d)$ and is therefore identified.[6]

This matrix factorization is not a Singular Value Decomposition because $\mathbf{F}_1, \mathbf{F}_2(d)$ are not orthogonal. This is why we need two matrices $P(0, d)$ and $P(1, d)$. Assumptions 2 and 3 guarantee that $\mathbf{F}_1, \mathbf{F}_2(d)$ and $D(0, d)$ are full rank. So $P(0, d)$ has rank H and the number of types is identified. To simplify, let us assume that $\mathbf{F}_1$ is a square matrix (the number of wages on the grid is chosen equal to H). Then,

$$P(1, d)P(0, d)^{-1} = \mathbf{F}_1 D(1, d) D(0, d)^{-1} \mathbf{F}_1^{-1}.$$

This last expression gives the eigendecomposition of the matrix $P(1, d)P(0, d)^{-1}$. Its eigenvalues are the elements of the diagonal matrix $D(1, d)D(0, d)^{-1} = \text{diag}\left[\frac{\pi(h,1,d)}{\pi(h,0,d)}\right]$.[7]

By Assumption 4, the ratios $\frac{\pi(h,1,d)}{\pi(h,0,d)}$ are all distinct, meaning that the eigenvalues of matrix $P(1, d)P(0, d)^{-1}$ are simple. It follows that the eigenvectors in $\mathbf{F}_1$ are identified up to scale. The unknown scale of the columns of $\mathbf{F}_1$ is identified because the first row of $\mathbf{F}_1$ is made of ones if the grid goes up to maximal wages. If the eigenvalues are not

[6] See Kasahara and Shimotsu (2014) on the specific identification of the number of groups.

[7] When $\mathbf{F}_1, \mathbf{F}_2(d)$ are square matrices we can identify them directly from the eigendecomposition of $P(1)P(0)^{-1}$, as we just did. However, using a finer grid making $\mathbf{F}_1, \mathbf{F}_2(d)$ rectangular may be useful to estimate the number of groups H. In the detailed proof in Appendix A, we use a finer grid making $\mathbf{F}_1, \mathbf{F}_2(d)$ rectangular and the Singular Value Decomposition of $P(0)$ to standardize (or "whiten") $P(1)$.

simple, then many choices are possible for the basis of the eigenspaces, and only some linear combinations of the mixture components $F_1(w \mid h)$ will be identifiable.

To sum up, the role of the instrument is to create two observable matrices $P(1,d)$ and $P(0,d)$ with the same algebraic structure. One is used to standardize the other (this is called "whitening" in the Independent Component Analysis literature), which gives to the ratios $\frac{\pi(h,1,d)}{\pi(h,0,d)}$ the interpretation of eigenvalues. Assumption 4 is also a condition for the point identification of the mixture components $F_1(w_1 \mid h)$. A symmetric argument proves the identification of $F_2(w_2 \mid d)$.

This proves identification given treatment d_i. Finally, how do we know that one group that we have labeled 1 for one particular value of d is the same as the group we have labeled 1 for another value? This is where Assumption 5 helps. The eigendecompositions for $d = 0$ and for $d = 1$ yield two values of $\mathbf{F}_1$ that should coincide. □

Our estimation method is described in Section 5 below. Although the identification proof is constructive, it leads to complex 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. 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 (flexible) parametric model after showing nonparametric identification is usual practice (see for example Cunha et al., 2010; Bonhomme et al., 2019).

4 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, emphasizing the conditions for their consistent estimation by OLS and IV. Some of the discussion here is well known. However, our definitions may differ slightly from what is usually considered in the literature. This is because our approach is conditional on unobserved heterogeneity.

Let $w_t(0)$ and $w_t(1)$ denote the *counterfactual outcomes*. Note that for pretreatment wages, $w_1(0) = w_1(1)$ by assumption. Let also $\Delta w_t(0)$ and $\Delta w_t(1)$ denote the wage changes between $t = 1$ and $t = 2, 3$ given training. Under the assumptions of our setup, counterfactual outcomes $w_t(0)$ and $w_t(1)$ satisfy the conditional independence assumption:

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

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

Define the observed outcome $w_t = d\, w_t(1) + (1-d)\, w_t(0)$ and similarly for Δw. We first derive the *Average Treatment Effect* (ATE) and the *Average Treatment Effect on the Treated* (ATT). Then, we consider Diff-in-diff OLS and IV estimators.

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

$$ATE_t(h) = \mathrm{E}\left[w_t(1) - w_t(0) \mid h\right] = \mu_t(h,1) - \mu_t(h,0),$$

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

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

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,

$$ATT_t(h) = \mathrm{E}\left[w_t(1) - w_t(0) \mid h, d = 1\right] = ATE_t(h).$$

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

$$ATT_t = \mathrm{E}\left[w_t(1) - w_t(0) \mid d = 1\right] = \sum_h \pi(h \mid d = 1) ATE_t(h), \tag{5}$$

with

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

Both ATE and ATT are identified under the structural model's assumptions.

DiD-OLS. Then, we study the OLS estimator of the impact of the treatment on the outcome change. The *difference-in-differences* (DiD) estimator is the OLS estimator: for $t = 2, 3$,

$$\begin{aligned} b_{OLS}(t) &= \frac{\mathrm{Cov}(\Delta w_t, d)}{\mathrm{Var}(d)} = \mathrm{E}\left[\Delta w_t(1) \mid d = 1\right] - \mathrm{E}\left[\Delta w_t(0) \mid d = 0\right] \\ &= \sum_h \pi(h \mid d = 1)\, \Delta\mu_t(h,1) - \sum_h \pi(h \mid d = 0)\, \Delta\mu_t(h,0) \\ &= ATT_t + B_{OLS}(t), \end{aligned}$$

where $B_{OLS}(t)$ is the bias, defined as

$$B_{OLS}(t) = \sum_h \left[\pi(h \mid d = 1) - \pi(h \mid d = 0)\right] \Delta\mu_t(h,0), \tag{6}$$

with $\Delta\mu_t(h,d) = \mu_t(h,d) - \mu_1(h)$.

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

1. $\pi(h \mid d=1) = \pi(h \mid d=0)$ for all types h; or

2. $\Delta\mu_t(h, d=0) = \Delta\mu_t(h=1, d=0)$ for all h, t.

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, Assumption 2 is the usual common trend assumption in DiD setups: the expected change in the outcome, for the untreated, is the same for everyone.

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 DiD-OLS estimator to be biased upward vis-a-vis the ATT. One can find a similar discussion in Carneiro et al. (2011).

DiD-IV. The IV estimator of the regression of Δw_t on d, using z as an instrument can be expressed as follows,

$$b_{IV}(t) = \frac{\mathrm{Cov}(\Delta w_t, z)}{\mathrm{Cov}(d, z)} = \frac{\mathrm{E}(\Delta w_t \mid z=1) - \mathrm{E}(\Delta w_t \mid z=0)}{\mathrm{E}(d \mid z=1) - \mathrm{E}(d \mid z=0)}.$$

The denominator of $b_{IV}(t)$ is trivially

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

The numerator can be factored as

$$\begin{aligned}
&\mathrm{E}(\Delta w_t \mid z=1) - \mathrm{E}(\Delta w_t \mid z=0) \\
&\qquad = \sum_h \left[\pi(h, d=1 \mid z=1)\, \Delta\mu_t(h,1) + \pi(h, d=0 \mid z=1)\, \Delta\mu_t(h,0)\right] \\
&\qquad - \sum_h \left[\pi(h, d=1 \mid z=0)\, \Delta\mu_t(h,1) + \pi(h, d=0 \mid z=0)\, \Delta\mu_t(h,0)\right] \\
&\qquad = \sum_h \left[\pi(h, d=1 \mid z=1) - \pi(h, d=1 \mid z=0)\right] \left[\mu_t(h,1) - \mu_t(h,0)\right] \\
&\qquad\qquad + \sum_h \left[\pi(h \mid z=1) - \pi(h \mid z=0)\right] \Delta\mu_t(h,0),
\end{aligned}$$

making use of

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

Hence,

$$b_{IV}(t) = LATE(t) + B_{IV}(t),$$

where we define

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

and

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

The LATE is a weighted average of conditional ATEs given type.[8] This average is informative if the weights are uniformly positive or negative, that is, if monotonicity holds (Imbens and Angrist, 1994):

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

with a strict inequality for at least one type. 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-differences, we can check whether all weights are of the same sign or not.

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

1. $\pi(h \mid z = 1) = \pi(h \mid z = 0)$ for all types h; or
2. $\Delta\mu_t(h, 0) = \Delta\mu_t(1, 0)$ for all h.

The second restriction has already been discussed in the case of OLS. The first restriction now links heterogeneity h to the instrument z instead of the treatment d. In our application, the instrument is defined at the worker and 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 regional 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).

[8] We call the parameter the LATE although our definition slightly differs from the original one in Imbens and Angrist (1994).

5 Application: the wage returns to training

In this section, we first present the data. We then perform some preliminary econometric analysis, estimating treatment effects on wage levels and first differences using OLS and 2SLS. Finally, we present the parametric specification and the maximum likelihood estimation method using the EM algorithm. We discuss the estimation of the number of groups H and how we construct confidence intervals (CIs) for the estimated parameters.

5.1 The data

We use survey data collected between 2013 and 2015 by Céreq[9] as part of the DEFIS survey.[10] 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.[11] 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. 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.[12] 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. We could keep only these workers (as we did in an earlier version of this paper). However, in doing

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

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

[11] 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

[12] 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.

so, we might lose some effect of training that might make a worker more employable. The estimation sample retains all workers, movers and stayers.

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.[13] 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.

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-resources 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 the effect of training on wages, in level and first differences, by OLS and IV.

5.2 OLS and IV

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).[14] This equation is first estimated

[13] 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.

[14] 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,

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

	All		Analysis	
	Trained	Untrained	Trained	Untrained
Demographics:				
Age (modal group)	40 to 44	45 to 49	40 to 44	45 to 49
Male	70.7	67.3	70.7	68.2
French	97.0	94.1	97.7	95.1
In couple	74.8	68.4	76.2	70.7
Has children	57.4	49.0	59.3	52.5
Disability	7.24	12.50	6.89	11.90
Previous health problem	3.39	5.72	3.01	5.26
Education:				
Less than high school diploma	28.3	46.1	27.4	45.0
High school diploma	18.5	18.6	18.3	18.6
Vocational degree	20.7	14.9	21.9	16.1
Bachelor's degree	7.93	5.48	7.87	5.60
Master's degree or more	23.9	13.8	23.9	13.8
Occupation:				
Unskilled production	5.88	9.58	5.58	9.05
Skilled production	18.5	26.2	17.3	25.4
Office worker	20.9	27.6	20.3	27.6
Foreman, supervisor	13.70	9.95	14.0	10.3
Technician, lower management	9.29	6.51	9.84	7.10
Engineer, manager	29.5	15.7	31.3	17.0
Job characteristics:				
Log(hourly wage), 2013 (w_1)	2.7	2.5	2.7	2.5
Log(hourly wage), 2014 (w_2)	2.8	2.6	2.7	2.6
Log(hourly wage), 2015 (w_3)	2.8	2.6	2.8	2.6
Permanent contract	90.0	83.3	93.7	91.4
Full time contract	88.7	80.1	91.1	85.3
Information on training (z)	78.8	62.8	80.2	65.3
Firm characteristics:				
3 to 49 employees	24.0	39.1	21.6	35.9
50 to 249 employees	20.5	21.8	20.6	22.8
250 to 499 employees	9.13	7.19	9.20	7.52
500 to 999 employees	8.58	6.53	8.73	6.84
1000 to 1999 employees	7.38	6.25	7.36	6.21
2000+ employees	30.4	19.1	32.5	20.7
Has HR department	89.6	81.5	90.8	83.1
Has individual incentive strategy	72.4	60.0	74.2	62.4
Has collective incentive strategy	78.4	64.5	80.5	67.7
Outsources part of activity	40.6	34.8	39.7	34.3
Number of observations	6343	9783	5110	6520

Notes: "All" refers to the whole sample and "Analysis" refers to the sub-sample of workers who remain in our analysis sample. 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 approximately 3,000 in 2014 and 2015.

Table 3: Static estimation of wage regressions with training

	OLS		2SLS	
	Without controls	With controls	Without controls	With controls
Log-wage levels				
2013	0.169	0.060	0.349	0.086
	(0.007)	(0.005)	(0.044)	(0.041)
2014	0.175	0.064	0.381	0.135
	(0.007)	(0.005)	(0.044)	(0.041)
2015	0.174	0.063	0.377	0.134
	(0.007)	(0.005)	(0.045)	(0.043)
Log-wage changes				
2014	0.005	0.004	0.032	0.049
	(0.003)	(0.003)	(0.019)	(0.027)
2015	0.004	0.003	0.029	0.048
	(0.004)	(0.004)	(0.023)	(0.031)
Nb of workers	11,628	11,628	11,628	11,628

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. Note first that the effect of training on pre-treatment wages remains significant even after adding many controls to the estimation. This is admittedly a lot less the case with IV and controls than with OLS.

The OLS-DiD results suggest very small effects of training (differences in the β's around 0.3-0.5% with and without controls). After instrumenting the training variable, we see stronger effects of around 3-5%. 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 5% with controls, which is non-negligible. The estimation of our structural model will help us understand better the nature of the differences between OLS and IV.

5.3 Parametric specification

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

existence of wage incentives for performance (individual and collective), whether the firm outsources activities. See Table 1 for summary statistics.

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,

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

and

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

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

The model is flexible at first and second order as long as parameters μ_t, σ_t are left unrestricted.

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

The data for each individual i is the array $x_i = (w_{i1}, w_{i2}, w_{i3}, z_i, d_i)$, $i = 1, ..., N$. 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

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

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

We use a sequential EM-algorithm for the likelihood maximization (see Appendix B for details). Moreover, we relabel groups h to be increasing in the value of $\mu_1(h)$.

5.4 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, but practical methods are rare, especially in situations where the correct model is not in the set of considered models (Fraley and Raftery, 1998). We use a range of criteria that we will describe in the results section.

5.5 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 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.[15] 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.

6 Results

We present the results in this section. We first explain how we choose the number of groups H. Next, we discuss the estimated distribution of groups, overall, and across values of instrument z and treatment d. Then, we discuss the empirical validity of Identification Assumption 4 and the validity of the common trend assumption. Finally, we compare the plugin estimates of treatment effects ATE, ATT and LATE with OLS and 2SLS.

6.1 Choosing the number of types H

The panels of Figure 1 present three criteria we use to choose the number of types for the remainder of our analysis. In Figure 1(a) the different broken lines show how total likelihood ($\ln L$) and penalized-likelihood criteria evolve with H. The two penalized-likelihood

[15] The validity of bootstrap for the estimation of mixtures by maximum likelihood rests on identifiability. Identifiability was discussed in Teicher (1963); Yakowitz and Spragins (1968). It takes the form of a rank condition such as Assumption 3 above. Then, in practice, one has to avoid group labeling to randomly change across bootstrap samples. There exist various techniques for that. Weighted-likelihood bootstrap is a popular one.

criteria are the well-known Akaike and Bayesian information criteria (respectively, AIC and BIC). We are looking for slope discontinuities ("elbows"), that is, values of H where the marginal gain in likelihood for an additional type is noticeably less than it is for $H-1$. There are elbows at $H = 3$ and then at $H = 7$ for AIC and BIC. The criteria are all then quite flat after $H = 8$.

In Figure 1(b) are displayed the smallest group sizes ($\min_h \pi(h)$) for each number of types. We do not want to let group sizes get too small and so studying this plot can help to select H. We can see that there are a number of points where the minimum group size drops off: around $H = 9$ and around $H = 19$.

Figure 1(c) plots the estimated autoregressive parameter ρ against H. The latent types should capture time-invariant heterogeneity across individuals, and hence we expect a ρ significantly smaller than one. The estimate of ρ drops sharply until $H = 8$ when it starts to level out, though it is still slightly decreasing until about $H = 20$.

Finally, in Figure 2, we plot the aggregate "treatment" effects, ATE. We first focus on the year 2013 preceding treatment. Although the model assumes $\mu_1(h)$ independent of d, nothing prevents the calculation of a mean wage in 2013 by type h and treatment d. This can be done in a post-convergence M-step. The counterfactual ATEs should be zero under our model assumptions. Thus, we want to select a number of types which gives an ATE in 2013 very close to zero. We see that ATE(2013) becomes truly negligible only for very large values of $H \geq 19$. At the same time, the treatment effects for 2014 and 2015 become larger for these large values of H. Thus, we conclude that only for a large H can we be reasonably sure that unobserved heterogeneity has exhausted most of the spurious effect of training and is able to reveal the full extent of the effect of the treatment.

Overall, $H = 19$ seems to be a conservative choice to focus on for the rest of this paper, although we have analyzed other slightly lower values that lead to similar results.

6.2 Marginal and conditional type distributions

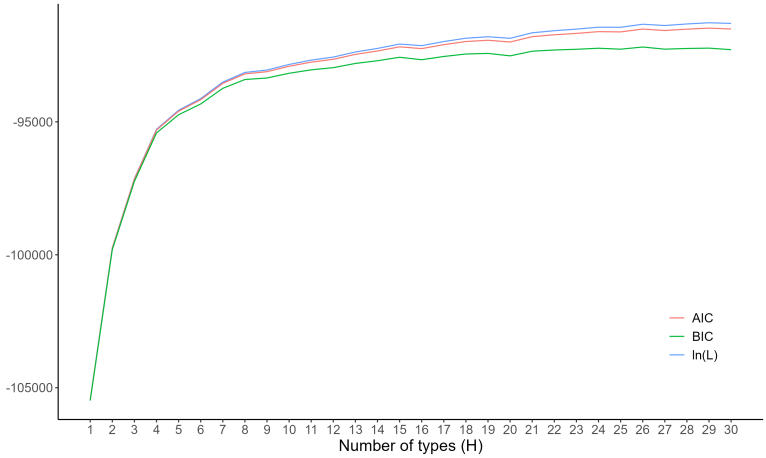
In Figure 3, we study the composition of the sample in terms of types $h \in \{1, ..., H = 19\}$. The error bars indicate bootstrapped, 90% confidence intervals. Group sizes $\pi(h)$ are displayed in Panel (a); the heterogeneity distribution by value of the instrument $\pi(h \mid z)$ and by value of the treatment $\pi(h \mid d)$ are shown in Panels (b) and (c).

Next, as emphasized in Section 4, the IV estimator is equal to the LATE if the instrument is well randomized, i.e. $\pi(h \mid z = 1) = \pi(h \mid z = 0)$, which would require the red and blue bars in Figure 3(b) to be equal for each h. This property appears to be violated here, with lower types being less likely to receive information on training (red bars larger than blue), while middle to high types are more likely to receive information (blue bars larger).

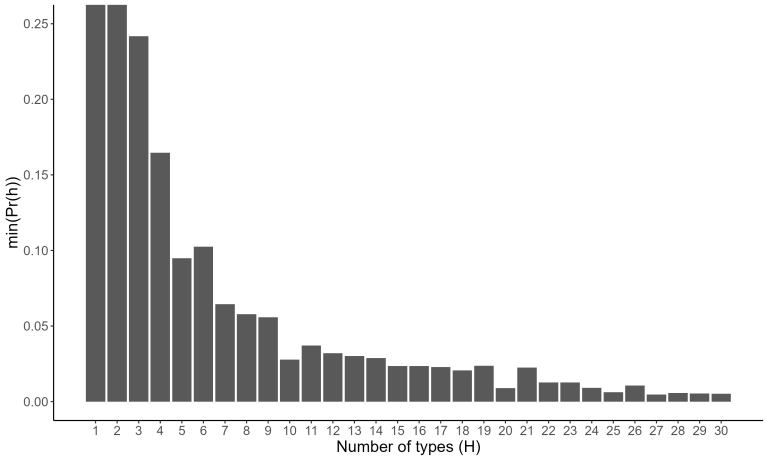
The opposition low type/high type is here even more pronounced for the distribution

Figure 1: Criteria for choosing the number of types H

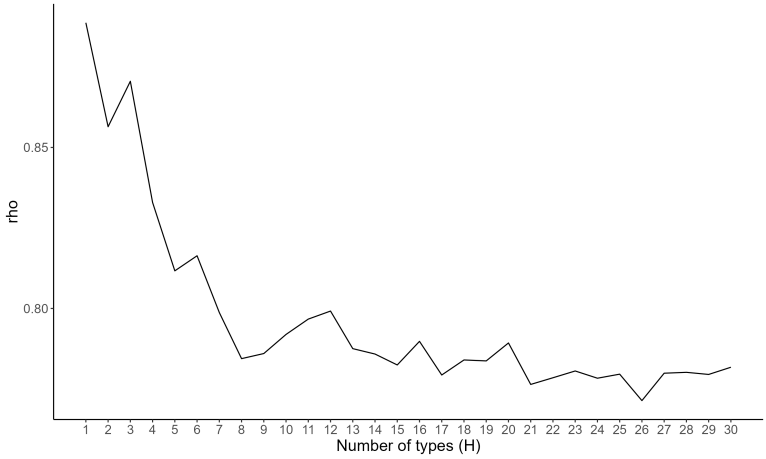
(a) Likelihood criteria



(b) Minimum group size $\pi(h)$ by H

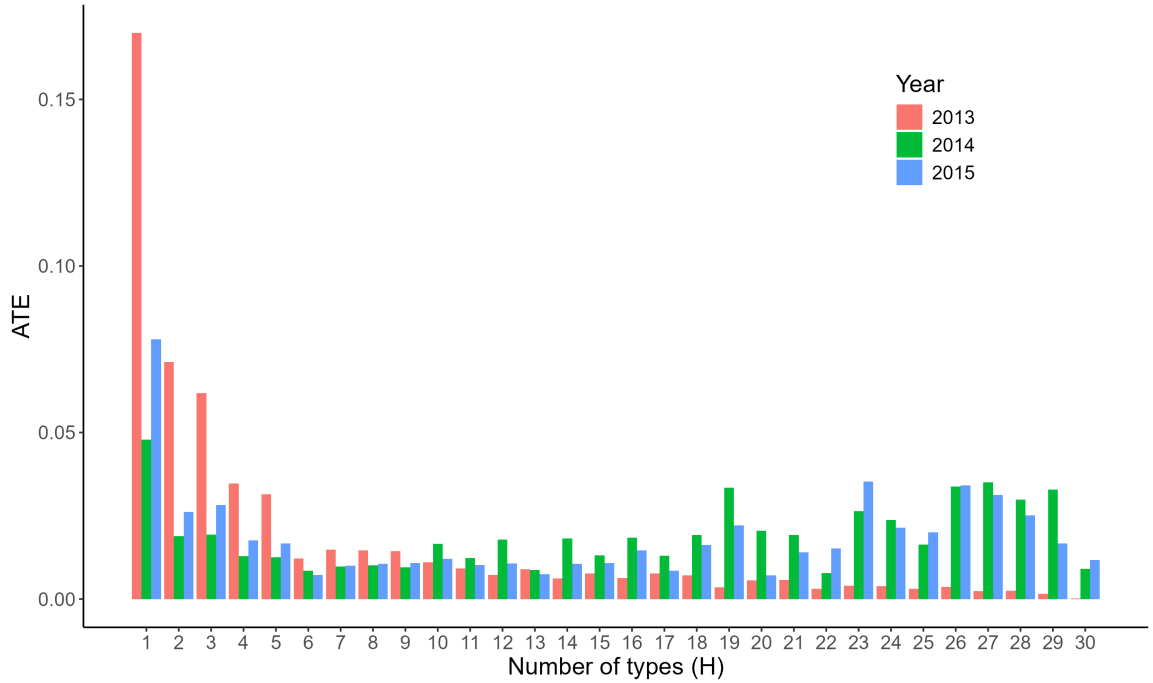


(c) Autoregressive parameter ρ



Notes: Likelihood criteria: 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 $-BIC$.

Figure 2: ATE for values of $H \leq 30$



Notes: The solid (transparent) bars are the estimated ATE (ATT) for a given H.

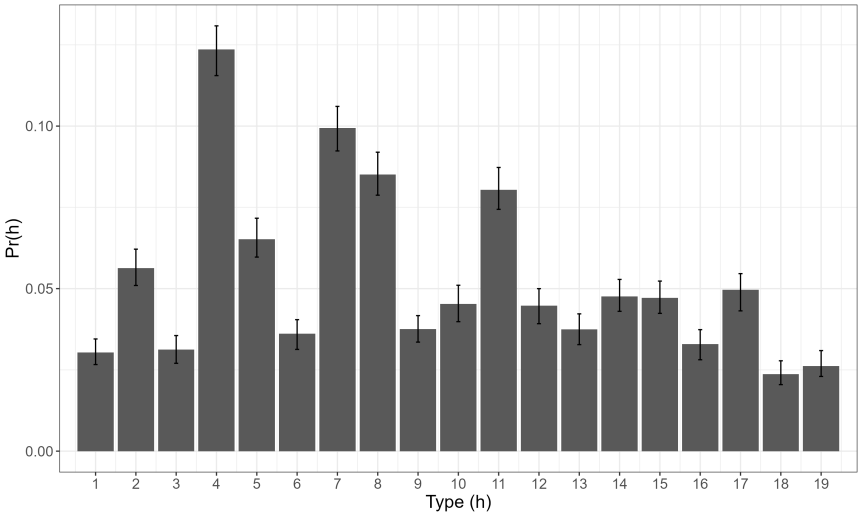
of types given treatment, which indicates a degree of compliance (see next section). It follows that the DiD-OLS and DiD-IV will be biased estimates of ATT and LATE unless the common trend assumption is satisfied.[16]

We end this section with an answer to the question: what are these types? The EM algorithm allows to estimate a posterior type probability for all workers. We can thus predict pretreatment wages for all workers as $\mu_{1i} = \sum_h p_i(h)\mu_1(h)$, where $p_i(h)$ is the posterior type probability of individual i being of type h and $\mu_1(h)$ is the mean pretreatment wage given type. This provides a continuous measure of the predicted type for all workers that we can correlate with observed characteristics. The regression output is displayed in Table 4. Higher types are older, more educated, employed in more skilled occupations and in larger firms. They are more often male, in a couple and in better health. Better types also tend to be matched with better jobs: open-ended, full-time contracts and with employers that provide individual or collective incentives to their employees. Controlling for all these characteristics does not exhaust the positive correlation between the latent type and training. All observed characteristics and training predict the type index μ_{1i} well, with an R-square close to 60%, but a very significant fraction of the variance remains unexplained. This is indicative that there is unobserved heterogeneity on top of all these observed controls and that the 19 latent types are able to capture multi-dimensional heterogeneity well.

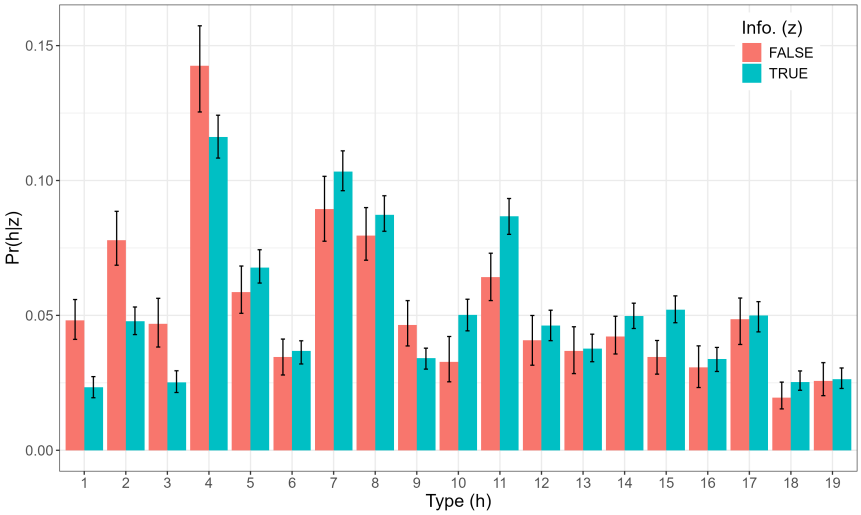
[16]Note that $\sum_h \pi(h \mid z) = \sum_h \pi(h \mid d) = 1$. A bigger proportion of some types in the treated group implies a smaller proportion of other types.

Figure 3: Type distributions

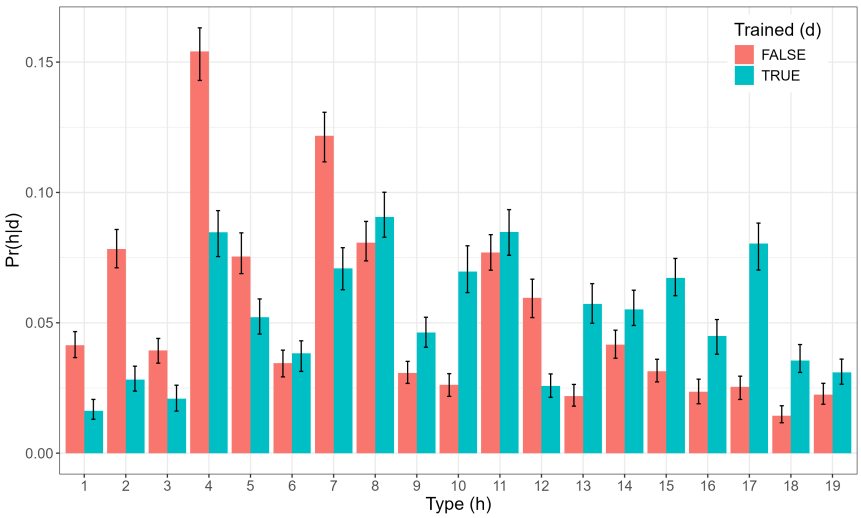
(a) Group sizes $\pi(h)$



(b) By instrument value, $\pi(h \mid z)$



(c) By treatment value, $\pi(h \mid d)$



Notes: Error bars show 90% confidence intervals, obtained by bootstrap.

Table 4: Link between observed and unobserved characteristics

Train	0.057^{***}	(0.004)	*Socio-professional status*		
			Unskilled production	-0.392^{***}	(0.010)
Age			Skilled production	-0.372^{***}	(0.008)
15 to 19 years	-0.139^{***}	(0.024)	Office worker	-0.369^{***}	(0.007)
20 to 24 years	-0.2010^{***}	(0.010)	Foreman, supervisor	-0.199^{***}	(0.008)
25 to 29 years	-0.165^{***}	(0.008)	Technician, lower manager	-0.282^{***}	(0.009)
30 to 34 years	-0.089^{***}	(0.008)	Engineer, executive	0	
35 to 39 years	-0.0340^{***}	(0.008)			
40 to 44 years	0		*Firm size*		
45 to 49 years	0.0275^{***}	(0.007)	3-9	0	
50 to 54 years	0.062^{***}	(0.008)	10-19	0.021^{**}	(0.011)
55 to 59 years	0.080^{***}	(0.009)	20-49	0.028^{***}	(0.011)
60 to 64 years	0.063^{***}	(0.018)	50-249	0.029^{***}	(0.011)
65 to 69 years	0.069	(0.035)	250-499	0.058^{***}	(0.011)
70 years and over	-0.149^{**}	(0.084)	500-999	0.073^{***}	(0.012)
			1000-1999	0.085^{***}	(0.013)
Education			>2000	0.100^{***}	(0.011)
Less than high school	-0.140^{***}	(0.009)			
High school	-0.068^{***}	(0.010)	*Job characteristics*		
Vocational	-0.007^{***}	(0.008)	Open ended contract	0.020^{*}	(0.009)
Bachelor	0		Full time	0.053^{***}	(0.007)
Master or more	0.065^{***}	(0.010)	HR department	0.008	(0.007)
			Individual incentives	0.012^{***}	(0.005)
Female	-0.077^{***}	(0.005)	Collective incentives	0.017^{***}	(0.006)
Couple	0.026^{***}	(0.005)	Outsourcing	0.020^{***}	(0.004)
Disability	-0.032^{***}	(0.007)			
Previous health issues	-0.042^{***}	(0.010)	Constant	2.785^{***}	(0.018)

Observations	11,222
R^2	0.580
Adjusted R^2	0.579
Residual Std. Error (df = 11183)	0.220
F Statistic (df = 38; 11183)	407.156^{***}

Notes: This table shows the regression of predicted wages in 2013, $\mu_{1i} = \sum_h p_i(h)\mu_1(h)$, where $p_i(h)$ is the posterior type probability of individual i being of type h and $\mu_1(h)$ is the mean pretreatment wage given type, on observed individual and job characteristics.

6.3 Compliance

Figure 4 investigates compliance. We begin to display, in Panel (a), the probabilities of being treated conditional on type and the instrument value, i.e., $\mathrm{E}(d \mid h, z) = \pi(d = 1 \mid h, z)$ for all $h \in \{1, ..., H = 19\}$ and $z \in \{0, 1\}$. 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 perfectly here: those who receive information on training are more likely to train across all types. Second, the bars are generally increasing by type: always taking and compliance seem to be increasing in unobserved type (commending higher wages). Thus, there is a pattern that training is more likely to be offered to more skilled types, and skilled workers are more likely to train even if they have not been informed about training possibilities by their employers.

Panel (b) displays the symmetric probability of being informed conditional on type and the treatment value, i.e., $\mathrm{E}(z \mid h, d) = \pi(z = 1 \mid h, d)$. Recall that Assumption 4 requires

$$\frac{\pi(h, 1, d)}{\pi(h, 0, d)} = \frac{\pi(z = 1 \mid h, d)}{1 - \pi(z = 1 \mid h, d)}$$

to vary with h for the structural model to be identified. Notice that

$$\frac{\pi(h, 1, d)}{\pi(h, 0, d)} = \frac{\pi(d \mid h, z = 1)\pi(h \mid z = 1)\pi(z = 1)}{\pi(d \mid h, z = 0)\pi(h \mid z = 0)\pi(z = 0)}.$$

We can see from Panel (a) that $\frac{\pi(d=1|h,z=1)}{\pi(d=1|h,z=0)}$ is decreasing in h, and from Figure 3(b) that $\frac{\pi(h|z=1)}{\pi(h|z=0)}$ is increasing. After $h = 10$, both ratios tend to stabilize. The resulting effect on $\mathrm{E}(z \mid h, d)$ or $\frac{\pi(h,1,d)}{\pi(h,0,d)}$ is a greater stability. We can see from Figure 4(b) that there is some variability of $\pi(z = 1 \mid h, d)$ before $h = 5$ and much stability above. We can therefore expect some difficulty in accurately identifying component distributions for larger values of h. Whether there is enough variation of $\pi(z = 1 \mid h, d)$ with h for the model to be identified should be seen in the width of Bootstrap confidence intervals.

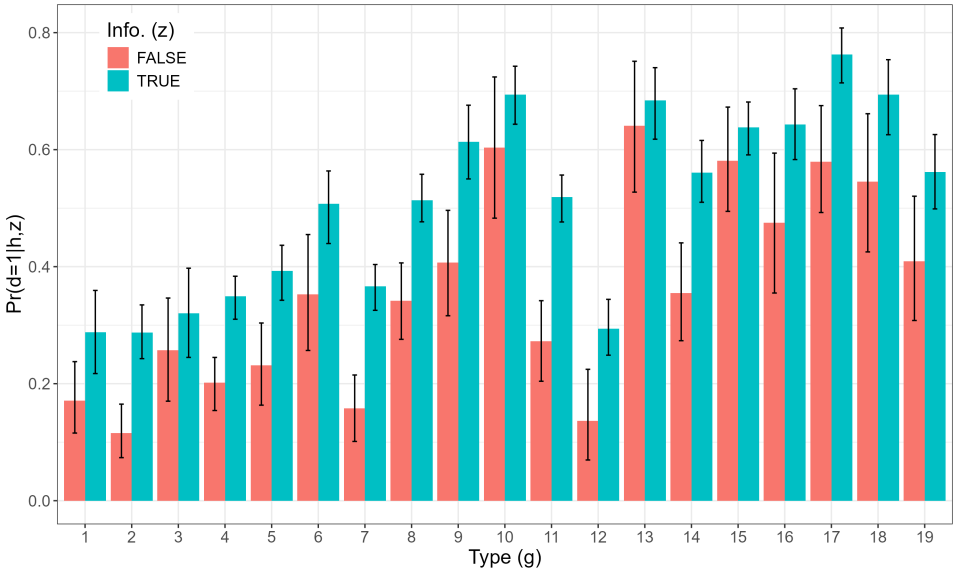
6.4 Conditional treatment effects

Figure 5 displays type-conditional average treatment effects $ATE(h)$ for all $h \in \{1, ..., H\}$. There is substantial heterogeneity in treatment effects across types, although the effects are generally small and positive. A few groups, though, exhibit large and significant effects, but they are not the largest ones.

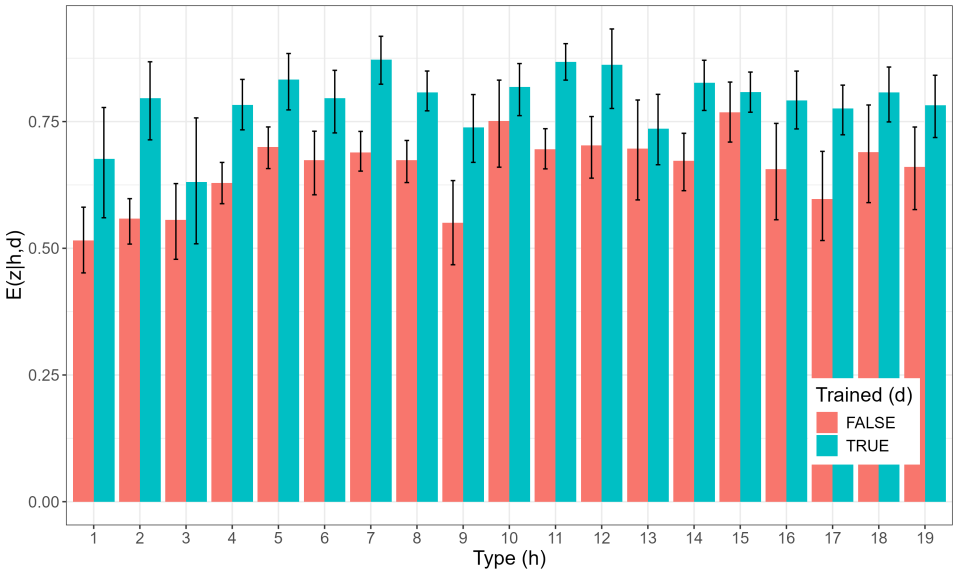
Note that we calculate an empirical wage mean in 2013 for the trained and the untrained, and corresponding pseudo-ATEs, as in Figure 2. Again, we did this calculation to verify the predetermination assumption *ex post*. The algorithm performs well along this dimension, with $ATE(h)$s for 2013 generally small relative to 2014 and 2015. Moreover, the 90% bootstrapped confidence intervals (black error bars) include zero for the majority

Figure 4: Treatment and instrument probability

(a) Treatment probability $\pi(d = 1 \mid h, z)$



(b) Rank condition, $\pi(z = 1 \mid h, d)$



Notes: Error bars show 90% confidence intervals, obtained by bootstrap.

of types.

Lastly, we show in Figure 6 a graphical test of the common trend assumption. We see some large fluctuations of $\mu_t(h, d = 0) - \mu_{2013}(h, d = 0)$, for $t = 2014, 15$, apparently significant, with both positive and negative values. This certainly explains why the DiD-OLS effect of training was estimated very close to zero.

6.5 Average treatment effects

Finally, we aggregate across types to obtain a variety of treatment effects summarizing the whole sample, which are presented in Table 5.

The first three columns are estimates of the ATE, ATT, and LATE. These are plug-in estimates from the structural model. We take the conditional mean wage estimates $\mu_t(h, d)$ in years $t = 2013, 2014, 2015$, and we calculate the ATE, ATT and LATE as in equations (4), (5) and (7).

Let then $b_{OLS} = ATT + B_{OLS}$ and $b_{IV} = LATE + B_{IV}$ denote the corresponding plug-in estimates of the OLS and IV parameters

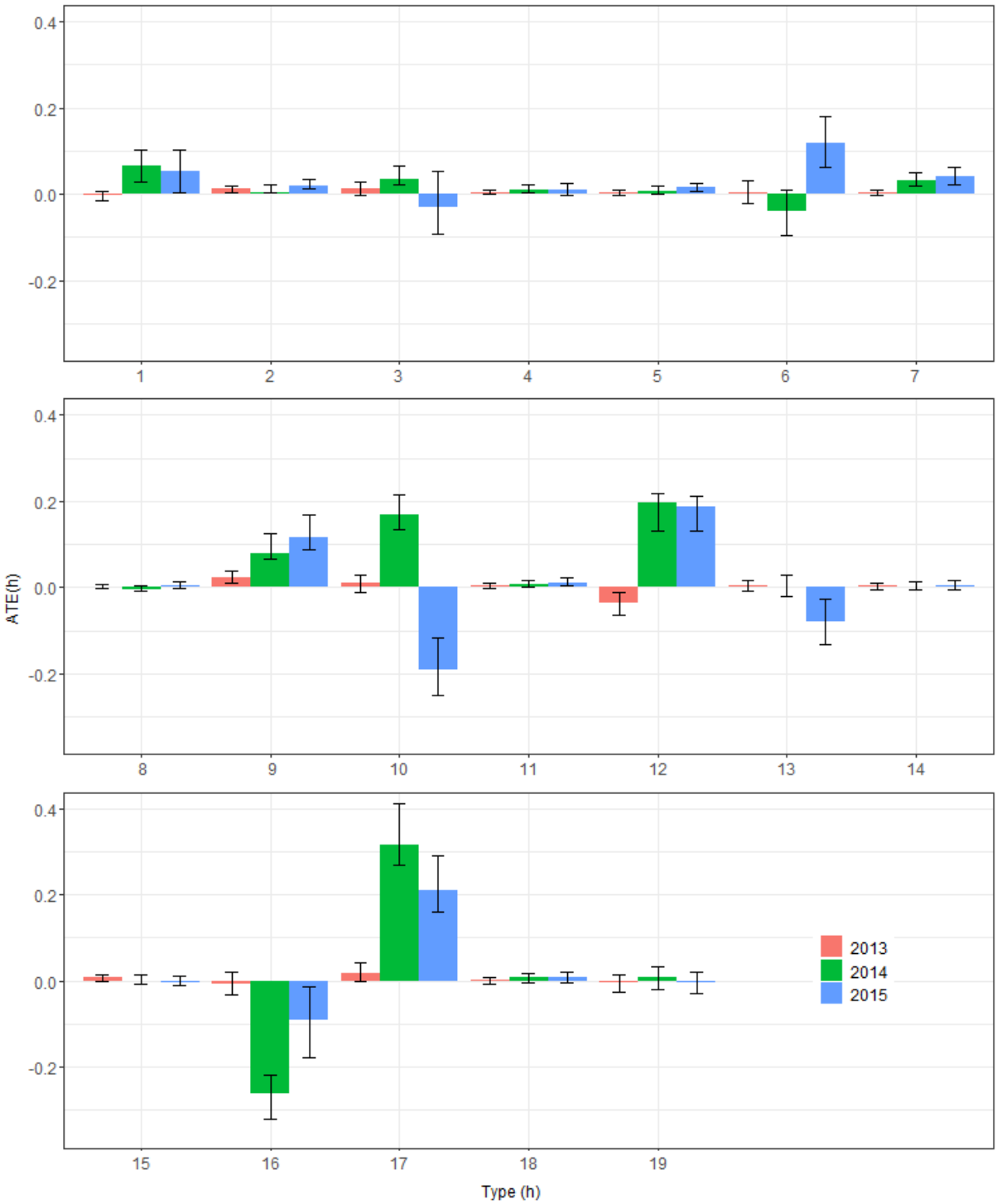
$$\frac{\text{Cov}(w_t, d)}{\text{Var}(d)} \quad \text{and} \quad \frac{\text{Cov}(w_t, z)}{\text{Cov}(d, z)} \tag{11}$$

for wage levels w_t in all three years, where B_{OLS} and B_{IV} are their biases calculated using equations (6) and (8). The biases are huge, making most of the value of the parameters. They arise because of heterogeneous distributions of the treatment d and the instrument z, and heterogeneous wage trends given treatment. The very large IV bias is indicative that the instrument is not perfectly randomized: $\pi(h \mid z = 1) \neq \pi(h \mid z = 0)$.

Finally, we denote as $\widehat{b}_{OLS}$ and $\widehat{b}_{IV}$ the standard, analog OLS and the IV estimators obtained by replacing the population variances in equation 11 by sample variances. While for OLS the plug-in and analog estimates coincide ($b_{OLS} = \widehat{b}_{OLS}$), for IV there may be an additional bias because, in the sample, pretreatment wages may be correlated with the treatment and the instrument, and post-treatment wages may be correlated with the instrument given the treatment (Assumptions 5 and 4). These two assumptions are necessary for our structural model to be identified, and are imposed in the Maximum Likelihood estimation. However, they are not imposed by the 2SLS estimator. The difference $\widehat{b}_{IV} - b_{IV}$ is displayed in column 9 of Table 5. Although the magnitudes of these biases are sizable in comparison to the treatment effect values, their Bootstrap standard errors are even larger. Hence, Assumptions 5 and 4 are not rejected.

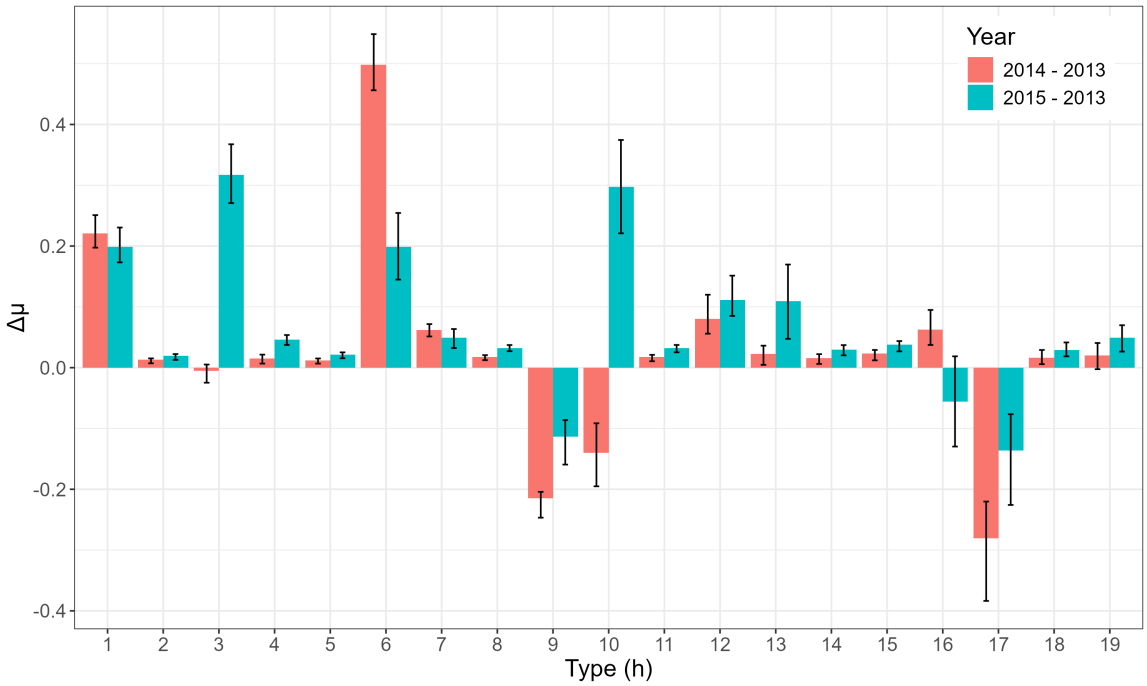
In the end, we find similar-sized estimates of the ATE, ATT, and LATE, of around 3-4% in 2014, falling to $1.5 - 2\%$ in 2015. These are quite precisely estimated. The DiD-2SLS estimator with and without controls is of comparable magnitude but a lot less precise. The DiD-OLS estimates seems to exhibit a severe attenuation bias.

Figure 5: Type-conditional treatment effects



Notes: Type-conditional average treatment effects: $ATE_t(h) = \mu_t(h, d = 1) - \mu_t(h, d = 0)$. The error bars are 90% confidence intervals obtained by bootstrap.

Figure 6: Test of common trends



Notes: The bars show the growth in (log)-wages of the untrained, $\Delta\mu_t(h, d = 0) = \mu_t(h, d = 0) - \mu_1(h)$, for 2014 ($t = 2$) and 2015 ($t = 3$). Under the common trend assumption, $\Delta\mu_t(h, d = 0)$ should be independent of h. The error bars represent 90% confidence intervals, obtained by bootstrap.

Table 5: Aggregate treatment effects

	ATE	ATT	LATE	$\widehat{b}_{OLS} = b_{OLS}$	B_{OLS}	$\widehat{b}_{IV}$	b_{IV}	B_{IV}	$\widehat{b}_{IV} - b_{IV}$
2013	.004	.005	.003	.169	.165	.348	.340	.337	.008
	(.002)	(.002)	(.002)	(.007)	(.007)	(.046)	(.046)	(.046)	(.010)
2014	.033	.038	.034	.173	.134	.380	.357	.323	.023
	(.004)	(.005)	(.014)	(.007)	(.008)	(.046)	(.045)	(.046)	(.012)
2015	.022	.016	.015	.172	.155	.377	.333	.318	.044
	(.004)	(.004)	(.011)	(.007)	(.008)	(.046)	(.043)	(.045)	(.016)

Notes: (1) Standard errors are in parentheses, calculated as the standard deviation of the parameter estimates from 500 weighted-likelihood bootstrap repetitions. (2) $\widehat{b}_{OLS}$ and $\widehat{b}_{IV}$ are "naive" estimates obtained using ordinary least squares (OLS) and two-stage least squares (IV). (3) b_{OLS} and b_{IV} are the plug-in estimates, calculated from structural estimates using the formulas in Section 4.

7 Conclusion

In this article, we developed and demonstrated the empirical use of discrete mixtures 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-differences 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 have unobserved heterogeneity conditioning 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 positive effect on wages, around 4% on average.

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.

Athey, S. and G. W. Imbens (2006): "Identification and inference in nonlinear difference-in-differences models," *Econometrica*, 74, 431–497.

——— (2017): "Chapter 3 - The Econometrics of Randomized Experiments," in *Handbook of Economic Field Experiments*, ed. by A. V. Banerjee and E. Duflo, North-Holland, vol. 1 of *Handbook of Field Experiments*, 73–140.

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].

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.

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A Distributional Framework for Matched Employer Employee Data," *Econometrica*, 87, 699–739.

Bonhomme, S. and J.-M. Robin (2009): "Consistent noisy independent component analysis," *Journal of Econometrics*, 149, 12–25.

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.

Hall, P. and X.-H. Zhou (2003): "Nonparametric estimation of component distributions in a multivariate mixture," *Annals of Statistics*, 31, 201–224.

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.

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

——— (2014): "Nonparametric identification and estimation of the number of components in multivariate mixtures," *Journal of the Royal Statistical Society, Series B*, 76, 97–111.

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.

Newey, W. K. (2013): "Nonparametric Instrumental Variables Estimation," *American Economic Review*, 103, 550–56.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

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.

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

Teicher, H. (1963): "Identifiability of Finite Mixtures," *The Annals of Mathematical Statistics*, 34, 1265 – 1269.

Yakowitz, S. J. and J. D. Spragins (1968): "On the Identifiability of Finite Mixtures," *The Annals of Mathematical Statistics*, 39, 209–214.

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 $f_2(w_2 \mid h, d) \neq 0$ at least for one h — we can write

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

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 $f_2(w_2 \mid h, d) \neq 0$.

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

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

be the diagonal matrix with $\pi(h, z, d)\, f_2(w_2 \mid h, d)$ in the hth diagonal entry, keeping only the values of h such that $f(w_2 \mid h, d) \neq 0$. Let also $G_1(d, w_2) = \left[F_{1|2}(w_1 \mid w_2, h, d)\right]_{w_1 \times h}$ denote the matrix of pre-treatment wage probabilities, with w_1 indexing rows and h indexing columns. Similarly, let $G_2(d, w_2) = \left[F_{3|2}(w_3 \mid w_2, h, d)\right]_{w_3 \times h}$ be the post-treatment matrix. Again, the values of h indexing columns are only those such that $f_2(w_2 \mid h, d) \neq 0$. 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 $f_2(w_2 \mid h, d) \neq 0$ 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 (Whitening)**.** *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 labeling 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 labeling 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 $f_2(w_2 \mid h, d)$ (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 $f_2(w_2 \mid h, d)$.

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

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

sumption 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

We assume that the distribution of log-wages in period t, denoted w_t, is Normal and depends on type h and treatment d. Wages are given by the following expressions,

$$w_1 = \mu_1(h) + u_1, \quad u_1 \sim N\left(0, \sigma_1^2(h)\right) \tag{12}$$

$$w_t = \mu_t(h,d) + u_t, \quad u_t \sim N\left(\rho u_{t-1}, \sigma_t^2(h,d)\right), \quad t = 2,3 \tag{13}$$

The complete individual likelihood is given by:

$$\ell_{ih}(\beta) = \pi(h, z_i, d_i)\, f_1(w_{i1} \mid h)\, f_{2|1}(w_{i2} \mid w_{i1}, h, d_i)\, f_{3|2}(w_{i3} \mid w_{i2}, h, d_i) \tag{14}$$

We run the following procedure to estimate the parameters, iterating between an E-step (in which we update the posterior probabilities) and an M-step (in which we choose parameters to maximize the likelihood given the posteriors from the E-step).

E-step. The posterior probability of worker i to be of type h given data (i.e., the conditional probability of h knowing i, also called *responsibility*), denoted p_{ih}, 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,

$$p_{ih}^{(m)} = \frac{\ell_{ih}(\beta^{(m)})}{\sum_h \ell_{ih}(\beta^{(m)})}. \tag{15}$$

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

1. For $t = 1$:

$$\mu_1^{(m)}(h) = \frac{\sum_i p_{ih}^{(m)} w_{i1}}{\sum_i p_{ih}^{(m)}}, \tag{16}$$

$$(\sigma_1^2)^{(m)}(h) = \frac{\sum_i p_{ih}^{(m)} (u_{i1h}^{(m)})^2}{\sum_i p_{ih}^{(m)}}, \tag{17}$$

with $u_{i1h}^{(m)} = w_{i1} - \mu_1^{(m)}(h)$.

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

$$\mu_t^{(m)}(h,d) = \frac{\sum\limits_{\{i:d_i=d\}} p_{ih}^{(m)} \left[w_{it} - \rho^{(m-1)} u_{i,t-1,hd}^{(m)} \right]}{\sum\limits_{\{i:d_i=d\}} p_{ih}^{(m)}}$$

$$(\sigma_t^2)^{(m)}(h,d) = \frac{\sum\limits_{\{i:d_i=d\}} p_{ih}^{(m)} \left[u_{ithd}^{(m)} - \rho^{(m-1)} u_{i,t-1,hd}^{(m)} \right]^2}{\sum\limits_{\{i:d_i=d\}} p_{ih}^{(m)}},$$

where $u_{ithd}^{(m)} = w_{it} - \mu_t^{(m)}(h,d), t = 2, 3$.
Note that $\mu_t(h,d)$ now depends on ρ for $t = 2, 3$ because we impose $\mu_1(h,0) = \mu_1(h,1) = \mu_1(h)$, i.e., treatment d has no effect on pre-treatment wages, conditional on type (h). If we relaxed this constraint, the estimator $\mu_t(h,d)$ 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,

$$\rho^{(m)} = \frac{\sum\limits_h \sum\limits_{d\in\{0,1\}} \sum\limits_{i\in I(d)} p_{ih}^{(m)} \left(\frac{u_{i1h}^{(m)} u_{i2hd}^{(m)}}{(\sigma_2^2)^{(m)}(h,d)} + \frac{u_{i2hd}^{(m)} u_{i3hd}^{(m)}}{(\sigma_3^2)^{(m)}(h,d)} \right)}{\sum\limits_h \sum\limits_{d\in\{0,1\}} \sum\limits_{i\in I(d)} p_{ih}^{(m)} \left(\frac{(u_{i1h}^{(m)})^2}{(\sigma_2^2)^{(m)}(h,d)} + \frac{(u_{i2hd}^{(m)})^2}{(\sigma_3^2)^{(m)}(h,d)} \right)}.$$

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

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