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

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

November 20, 2020

Abstract

Using detailed survey data collected by Céreq (the Defis survey), linked to French administrative data on wages (DADS), we first carry out reduced-form analysis in the shape of difference-in-differences (DiD) estimates of the impact of formal training on wages and other outcomes. The estimated effects on wages are negligible, but we find small positive effects of training on the probability of gaining a permanent contract (CDI) and on becoming a full-time employee. These effects appear to be driven mainly by individuals who move firm. We then exploit a policy-relevant instrumental variable (IV)—whether the individual reports having received information on training from their firm. IV estimates of the same effects are larger in magnitude but are not statistically significant at even the 10% level. We then introduce unobserved heterogeneity as latent types. Our novel identification strategy allows us to use a binary instrument to estimate the effects of multivalued treatment. We estimate our model on the Defis data via the EM algorithm, and calculate analogues of the reduced-form treatment effects, to enable comparison with DiD and IV approaches. Our flexible model and estimation approach allows us to produce wage distributions conditional on worker type and training status.

Keywords: Heterogeneity; Wage distributions; Training; Matched employer-employee data

JEL codes: E24; E32; J63; J64

*Sciences Po, Paris; E-mail: oliver.cassagneaufrancis@sciencespo.fr

†CREST, ENSAE, Paris. E-mail: robert.gary-bobo@ensae.fr

‡University of Chicago; E-mail: julie.pernaudet@gmail.com

§Sciences Po, Paris and University College London. Email jeanmarc.robin@sciencespo.fr

1 Introduction

We are not heading towards Keynes' vision of a 15-hour work week; rather an increasingly cheaper supply of automated and offshore labour is changing the nature of the skills demanded of the labour force. Many displaced workers are not willing to suffer the loss in income that a return to formal education would require, making the ability to retrain while working a necessity. Given the current rate of technological progress, learning while earning is only likely to increase in importance. Governments are starting to catch on. Singapore recently introduced "lifelong learning accounts", providing its citizens with credits to spend on approved courses (Anonymous, 2017). France introduced a similar system in 2015, *le compte personnel de formation* (CPF), which entitles all employees to a certain number of hours training per year, funded by their employer and/or the state.[1] Skills also play an increasing role in inequality. The increases in wages paid to those at the top of the skill distribution while others' wages stagnate and fall in real terms suggest that recent increases in inequality are linked to skills (Autor et al., 2003). This is on top of the traditional view that career wage-paths are driven by the accumulation of human capital—through training as well as learning by doing. Ensuring the supply of skills keeps up with demand will require the ability to train while working for those already in the labour force, and this ability will likely play a vital role in the success of those yet to enter.

The apparent increasing importance of vocational training, and of an increasingly flexible workforce, raises important policy issues. It will be up to governments not only to intervene to address market failures in providing an efficient level of training (where economic theory has and will continue to play an important role), but also to ensure firms and workers have the evidence they need to make the right choices. This paper addresses the second concern, joining a growing literature which makes use of high-quality microdata to measure the returns to formal training undertaken during employment—as opposed to a related literature which focuses on training programmes for the unemployed. We make use of a novel French dataset collected by the *Centre d'études et des recherches sur les qualifications*[2] (Céreq), to estimate initially the wage returns to vocational training in France, though there is scope within both our model and data to investigate other returns. The estimated returns using OLS and differences-in-differences specifications are similar to those found in previous studies employing these specifications (both in France and in other countries). However, a 2SLS specification using a novel policy-relevant instrumental variable—whether the firm provides information on training to its employees—results in a much larger effect.

We then posit a model that allows for unobserved heterogeneity, and which relies on

[1] In fact, the French state has intervened in private sector training since the 1970s.

[2] Centre for studies and research into qualifications.

weaker exogeneity restrictions than a standard instrumental variable for identification. We estimate the model using the EM algorithm. A strength of this model and method is that we can estimate average treatment effects (ATEs) of training conditional both on unobserved type and training duration. We see some interesting variation in ATEs across types and duration: the ATE of training is often negative for the highest type (by mean pre-training wage), and the effect of different durations also varies across types. The next step in our analysis is to include controls in the model. It would also be interesting to look at outcomes other than wages.

The paper is organised as follows. The following section reviews the current state of the literature on training. Section 3 describes the dataset and presents our initial reduced-form analysis. Section 4 introduces the model underlying the structural analysis and discusses identification, with section 5 detailing our approach to estimation. We present some common *treatment effect* estimators in the framework of our model in section 6 and the results of our estimation via the EM algorithm are in section 7.

Literature review

The literature on the effect of training (and active labor market programs) is huge. The estimated impacts of training on wages and productivity are generally found positive; the effects on the risk of unemployment are often ambiguous. Several authors have reviewed this literature (see Heckman et al., 1999; Ferracci, 2013; McCall et al., 2016, and the meta-analyses of Card et al. (2010, 2018); Haelermans and Borghans (2012)).

Many of the most classic papers are based on non-experimental data with a panel structure and rely on fixed-effects estimators of the impact of training. 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. (1996), 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) have proposed an instrumental variable method to estimate the impact of treatments on quantiles, generalizing quantile regression, and applied their method to training data.[3] Other contributions control for selection in training using Heckman's two-stage estimator (*e.g.* LaLonde, 1986; Booth, 1993; Goux and Maurin, 2000).

A number of recent papers use randomized trials; see *e.g.* Attanasio et al. (2011); Grip and Sauermann (2012); Ba et al. (2017). The importance of how the comparison group is constructed is illustrated by Leuven and Oosterbeek (2008). They change their com-

[3] A behavioral approach to training participation is explored in Caliendo et al. (2016).

parison group (*i.e.* narrow it down) to pick only "workers who were willing to undertake training and whose employers were prepared to provide it, but did not attend the training course they wanted, due to some random event". This change of comparison group can reduce the estimated coefficient on training to almost zero.

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, that 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 is willing to subsidize training, or to share the benefits of training with the worker.[4] We should expect to see a positive effect of training on both productivity and wages. A number of papers use wage equations and production functions to test this prediction and do indeed find positive effects on productivity and wages.[5] Other contributions use matching estimators.[6]

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; Crépon et al. (2009), using methods developed in Abbring and van den Berg (2003).

Finally, an important question is to assess the importance and effects of unobserved heterogeneity, as well as the dynamic structure of the treatment effects of training (for a recent contribution proposing progress on these two fronts, see Rodríguez et al., 2018).

2 The Data

We use survey data collected between 2013 and 2015 by Céreq, as part of the DEFIS survey (*Dispositif d'enquêtes sur les formations et itinéraires des salariés*). The survey sampled 3,750 firms with three employees or more from all sectors but agriculture in 2013, and 16,126 workers were subsequently drawn from those firms' employees in the Fall of 2013. 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 are to be conducted. The first wave that we use in this paper interviewed employees between June and October 2015 on 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?").[7]

[4]See Acemoglu and Pischke (1998, 1999).

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

[6]see, *e.g.*, Brodaty et al. (2001); Gerfin and Lechner (2002); Kluve et al. (2012)

[7]The survey questions are in French; these are translations by the authors.

The responses to the employer survey (in 2013) and the worker survey (in 2015) are complemented by wage data obtained from tax registers (*Déclarations annuelles de salaires*, DADS) for the ongoing employment spells in December 2013 (13% in November), December 2014 and December 2015.[8] Our definition of the wage is 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. A fraction 78,1% (12,597/16,126) of workers remain employed by the same firm as in 2013 at the time of the interview in 2015. A greater fraction (89.2% = 12,100/13,562 and 85.3% = 11,103/13,014) of the wages recorded for 2014 and 2015 were paid by the same employer who paid the wage recorded in 2013. All this to say that attrition and job mobility can and will be neglected in this study.

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.[9] Table 1 presents the average characteristics of trained and untrained workers in terms of demographics, education, occupation and job characteristics, before any training (situation in the fall of 2013). All variables in rows are binary, except the 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, live as a couple, and 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. Using the employer survey (not presented here), we also find that trained workers are on average in bigger firms, that are more likely to have human resource staff and offer jobs that are more secure (i.e., a smaller proportion of workers under temporary or part-time contracts). Overall, if more advantaged workers are more likely to get training, then simply comparing the labor market outcomes of trained and untrained workers does not allow to recover the causal effect of training. The two right-hand columns focus on the subsample that we use in the majority of our analysis. This subsample excludes some individuals with "extreme" wage observations and more importantly includes only "stayers", workers who were at the same firm in 2015 as they were in 2013. The two subsamples 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

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

[9]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.

contract.

Table 1: Comparison of trained and untrained workers according to baseline characteristics

Subsample:	All		Analysis	
Variable (means):	Trained	Untrained	Trained	Untrained
Demographics:				
Age (modal group)	40-44	45-49	40-44	45-49
Male	70.6	67.3	71.4	68.9
French	93.3	89.9	97.9	95.1
In couple	74.8	68.4	77.2	72.3
Has children	57.4	49.0	61.5	54.6
Education:				
Less than high school diploma	28.3	46.1	26.7	46.1
High school diploma	18.5	18.6	18.4	18.1
Trade or vocational degree	20.7	14.9	21.8	16.3
Bachelor's degree	8.0	5.4	7.6	4.8
Master's degree or more	23.9	13.8	24.0	13.6
Occupation:				
Unskilled blue collar	5.1	9.1	4.4	8.5
Skilled worker, technician	18.4	26.2	17.5	25.1
Office worker, public sector employee	20.9	27.6	19.4	27.2
Foreman/Supervisor	11.2	7.4	12.2	8.3
Technician, draftsman, salesman	9.3	6.5	10.0	7.4
Engineer, manager	29.5	15.7	31.9	17.5
CEO, executive	2.5	2.6	2.2	2.4
Job characteristics:				
Log(hourly wage), 2013	2.74	2.56	2.74	2.56
Log(hourly wage), 2014	2.77	2.58	2.77	2.58
Log(hourly wage), 2015	2.79	2.61	2.80	2.61
Permanent contract	90.0	83.4	95.9	94.9
Full time contract	88.7	80.1	92.7	87.5
Number of observations	6,343	9,783	4,736	5,764

Notes: Mean trained gives the mean of each variable for workers who had training in 2014 or 2015, Mean untrained gives the mean for workers who had no training during those years. For all binary variables, the mean is given as a percentage. The bottom row gives the number of workers for all variables except log(hourly wage) for *all*, where approx. 50 observations are missing wages in 2013, and approx. 3,000 in 2014 and 2015.

In the next section, we attempt to address this selection issue using quasi-experimental reduced-form methods and study the causal effect of training on wage, as well as on the probability to hold a permanent or full-time job.

3 Reduced-form analysis

3.1 The effect of being trained

Difference-in-Differences. We start by exploiting the panel structure of the data in a difference-in-differences model. A worker is treated if they were trained between January 2014 and the time of the interview in 2015.

We also study the impact of training on two other outcomes from the employee survey: the probabilities of holding a permanent contract and of holding a full-time contract at the time of the survey. Workers are asked about their job in December 2013, and their current job, which allows us to observe their contract before and after treatment. For those two qualitative outcomes, and for the rest of the reduced-form analysis, we keep the same definition of treatment as for the wage outcome. Since the earliest declared training programs took place in January 2014, i.e., after the first observations (December 2013), and the second recorded point is 2015, i.e., after the latest 2014 wage observation, the difference-in-differences framework still applies: we observe the worker's contract before and after training. In the regressions, we control for workers who received training after their latest 2014 wage observation but before the 2015 survey. For all outcomes, we estimate the following difference-in-differences equation:

$$Y_{it} = \alpha + \beta A_t + \gamma D_i + \delta A_t.D_i + controls + u_{it}, \tag{1}$$

where Y_{it} is the outcome of worker i at time t ($t \in \{0,1\}$, before or after training), A_t is a binary variable taking value 1 after training, 0 otherwise, and D_i is a binary variable taking value 1 if worker i is treated (see the definition above), 0 otherwise. Controls include the workers' baseline characteristics presented in Table 1 (demographics, education, occupation and job characteristics), as well as health characteristics, previous employment and unemployment spells, and training programs received after the latest wage observation. We also control for important characteristics of the firm, such as its size, the percentage of temporary and part-time contracts, whether it has incentive-based pay system, whether part of the activities are outsourced, and whether it has a human resources deaprtment. These covariates help make the identification assumption more credible: we assume that keeping the observed characteristics fixed, treated workers would have on average the same outcome evolution as untreated workers, had they not received any training (*common trends* assumption). But the results are robust to variations in the choice of controls, including no covariates at all.

The estimates of equation (1) are presented in Table 2: for each of the four outcomes we study, each column displays all coeffficents except those of the controls, with standard errors in parentheses below. Apart from when moving is the dependent variable, each regression is run on the stayers and movers of our "analysis" sample separately. The

Table 2: Impact of training, DiD estimates (stayers vs movers)

Dependent variable:	Same firm	Log-wage		Permanent contract		Full-time	
Subsample:	All	Stayers	Movers	Stayers	Movers	Stayers	Movers
Year							
2014		0.03***	0.07***				
		(0.00)	(0.01)				
2015	-0.15***	0.06***	0.12***	0.01***	-0.26***	-0.00	-0.25***
	(0.00)	(0.00)	(0.01)	(0.00)	(0.02)	(0.01)	(0.02)
Trained	-0.01*	0.04***	0.01	-0.00	-0.01	0.01*	-0.03
	(0.00)	(0.01)	(0.02)	(0.00)	(0.03)	(0.01)	(0.02)
Interactions							
Trained × 2014		0.01	0.01				
		(0.01)	(0.02)				
Trained × 2015	0.05***	0.00	0.02	0.01	0.15***	0.01	0.18***
	(0.01)	(0.01)	(0.02)	(0.00)	(0.03)	(0.01)	(0.03)
Constant	0.74***	2.19***	2.35***	0.77***	0.31**	0.51***	0.17
	(0.03)	(0.03)	(0.04)	(0.03)	(0.11)	(0.04)	(0.10)
R^2	0.14	0.58	0.52	0.17	0.23	0.23	0.25
Adj. R^2	0.14	0.58	0.52	0.17	0.22	0.23	0.24
Num. obs.	22,860	30,018	4,272	20,012	2,848	20,012	2,848
RMSE	0.22	0.26	0.28	0.17	0.44	0.26	0.43

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Standard errors are in parentheses below estimates. All regressions include controls, for which a comprehensive list can be found in the Appendix. The sample is the same as the sample used to estimate the structural model. Workers with wages below the 1st or above the 99th percentiles are dropped.

results indicate that the impact of training on wages, at both dates, and across stayers and movers, is neglible. Similarly, the effects of training on permanent and full-time contracts are minimal for stayers. For movers, however, participation in at least one training program increases the probability of a permanent contract by 15 percentage points on average, and the probability of a full-time contract by 18 percentage points. Note that movers appear generally to do significantly worse than stayers in these categories; the coefficents on the 2015 year dummy are -0.26 and -0.25 for permanent and full-time contracts. Overall, our results suggest that training has the potential to reduce job insecurity.

Instrumental-variable estimation. We now try to identify the causal effect of training by regressing post-treatment outcomes on treatment status using an instrumental-variable (IV) estimator. Specifically, we instrument training by a binary variable indicating whether or not the employee reports receiving information on training at their firm.[10] Indeed, workers reporting that their firm "advertises" training have a higher probability of receiving training, by approximately 3 percentage points (t-stat of about 4). The exogeneity and exclusion restrictions cannot be tested, but we control for a large set of potential confounding factors, both at the employee and firm levels. If, nonetheless, advertising tends to induce the most able workers to invest in training, and if the observed characteristics that we include in the regressions fail to capture this variation in ability, then our instrument is invalid.

Table 3 presents the conditional LATE, for the same outcomes as before. No effect of training is estimated significantly different from zero. Several possible reasons can explain this discrepancy with the previous results. Note, however, that standard errors are large and the estimated effects are sizable.

3.2 Training intensity: duration

In this section, we study whether the impact of training varies with training duration. Previously, we only considered binary treatment, trained or not, but an important dimension of heterogeneity in training is duration. Workers can receive training over vastly different durations, from very short training sessions (often mandatory training; e.g. workplace health and safety procedures), to many-month-long training programs that are more likely to target skill formation. The total number of hours of training workers received varies broadly in our sample, with peaks corresponding to 6 hours (1 day), 12 hours (2 days), 18 hours (3 days), and so on (see figure 1). There does not seem to be any problematic correlation between the observations we exclude from our analysis and training duration; the relative size of the top sections of the bars in figure 1 appears stable.

[10] We have two similar variables in our data; one is this worker-reported measure, and a second is a firm-reported measure of whether the firm actively supplies information on training to their workers. Which of these variables to use was a source of discussion; one is possibly contaminated by selection at the individual level, and the other at the firm level. Our analysis of the correlation of observables and possible instruments suggests that the required exogeneity assumption is better supported by the worker-reported variable.

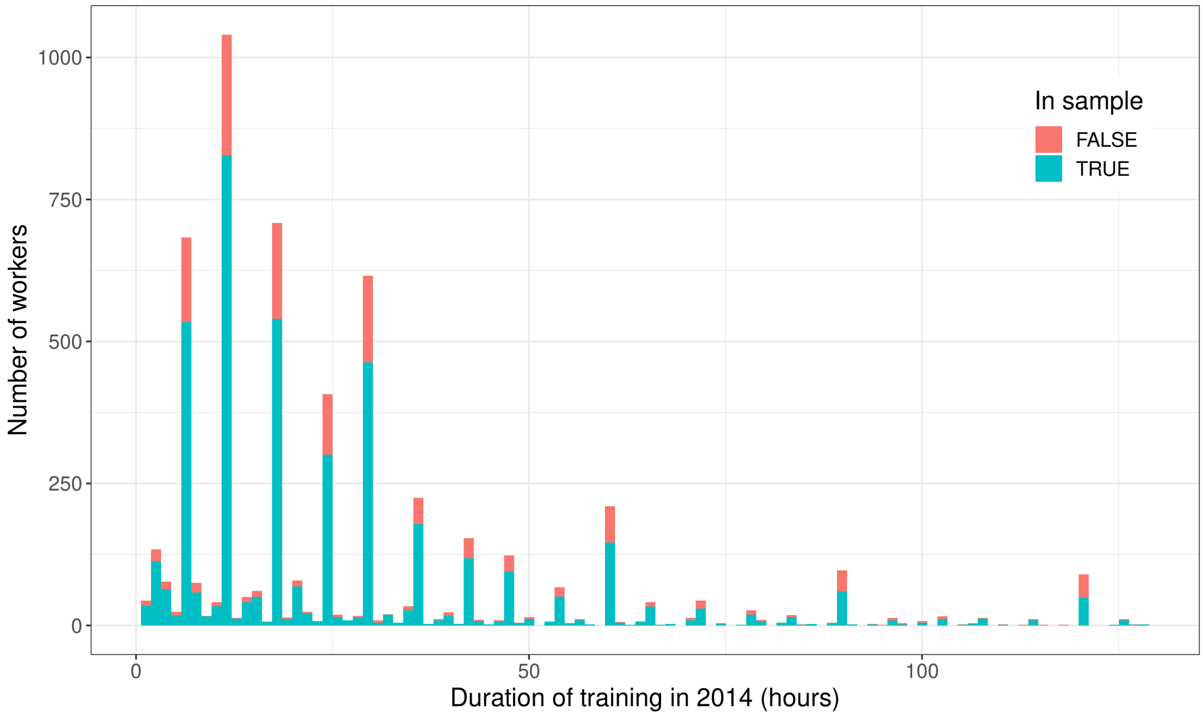
Table 3: Impact of training, IV estimates (stayers vs movers)

Dependent variable:	Same firm	Log-wage		Permanent contract		Full-time	
Subsample:	All	Stayers	Movers	Stayers	Movers	Stayers	Movers
Year							
2014		0.01	0.03				
		(0.02)	(0.06)				
2015	-0.32***	0.04	0.03	0.01	-0.18	-0.00	-0.27**
	(0.02)	(0.02)	(0.06)	(0.01)	(0.10)	(0.02)	(0.09)
Trained	-0.04	0.09*	0.00	0.08**	0.55*	0.19***	0.09
	(0.04)	(0.04)	(0.13)	(0.03)	(0.22)	(0.04)	(0.20)
Interactions							
Trained × 2014		0.05	0.12				
		(0.05)	(0.17)				
Trained × 2015	0.45***	0.04	0.28	0.01	-0.08	0.01	0.23
	(0.04)	(0.05)	(0.17)	(0.03)	(0.28)	(0.05)	(0.25)
Constant	0.90***	2.22***	2.38***	0.81***	0.27*	0.58***	0.18
	(0.04)	(0.03)	(0.05)	(0.03)	(0.13)	(0.05)	(0.11)
R^2	-0.13	0.57	0.49	0.13	0.06	0.15	0.23
Adj. R^2	-0.14	0.57	0.48	0.13	0.04	0.15	0.22
Num. obs.	22860	30018	4272	20012	2848	20012	2848
RMSE	0.26	0.26	0.29	0.18	0.49	0.28	0.43

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Standard errors are in parentheses below estimates. All regressions include controls, for which a comprehensive list can be found in the Appendix. The sample is the same as the sample used to estimate the structural model. Workers with wages below the 1st or above the 99th percentiles are dropped.

Figure 1: Distribution of training durations



We now define the treatment variable d_i as taking five different values: no training; up to 2 days; 2 to 5 days; 1 to 4 weeks; and more than a month. As in the previous section, we estimate first the effects under the DiD specification, and then turn to IV estimation using information on training as our instrument. Note that this is *total duration*, so a value of 4 months could encompass a single long training episode, or multiple shorter episodes.

DiD estimates. Table 4 presents for each of the four types of duration and four outcomes the DiD estimate with its standard error and the p-value. Here again, the impacts on wage are not significantly different from zero, whatever the duration, at both dates, though the coefficients are larger in magnitude for movers. When permanent and full-time contracts are the dependent variables, we see negligible effects for stayers, and much larger (and statistically significant) effects for all movers but those receiving training for over 4 weeks. The effects on contract type are between 13*p.p.* (shortest training on the probability of a permanent contract) and 25*p.p.* (1-4 weeks of training on the probability of a full-time contract). Note that these effects need to be viewed alongside the coefficients in the second row; movers are 25*p.p.* *less likely* to hold a permanent or full-time contract in 2015 than in 2013.

Table 4: Impact of training as a function of duration, DiD estimates (stayer vs movers)

Dependent variable:	Same firm	Log-wage		Permanent contract		Full-time	
Subsample:	All	Stayers	Movers	Stayers	Movers	Stayers	Movers
Year							
2014		0.03***	0.07***				
		(0.00)	(0.01)				
2015	-0.15***	0.06***	0.12***	0.01***	-0.26***	-0.00	-0.25***
	(0.00)	(0.00)	(0.01)	(0.00)	(0.02)	(0.01)	(0.02)
Training duration							
]0, 2] days	-0.01	0.03***	0.01	-0.00	-0.00	0.00	-0.01
	(0.01)	(0.01)	(0.03)	(0.00)	(0.04)	(0.01)	(0.04)
]2, 5] days	-0.02*	0.04***	-0.01	0.01	-0.00	0.03**	-0.02
	(0.01)	(0.01)	(0.03)	(0.01)	(0.04)	(0.01)	(0.04)
]1, 4] weeks	-0.01	0.06***	0.02	-0.01	-0.02	0.02*	-0.06
	(0.01)	(0.01)	(0.03)	(0.01)	(0.04)	(0.01)	(0.04)
>4 weeks	0.00	0.06***	0.03	-0.02	-0.02	-0.01	-0.04
	(0.01)	(0.01)	(0.03)	(0.01)	(0.05)	(0.01)	(0.05)
Duration × *Year*							
$]0,2]$ days × 2014		0.00	0.02				
		(0.01)	(0.04)				
$]2,5]$ days × 2014		0.00	0.03				
		(0.01)	(0.04)				
$]1,4]$ weeks × 2014		0.01	0.03				
		(0.01)	(0.04)				
>4 weeks × 2014		0.01	-0.04				
		(0.02)	(0.04)				
$]0,2]$ days × 2015	0.08***	-0.00	0.05	0.01	0.13*	0.01	0.19**
	(0.01)	(0.01)	(0.04)	(0.01)	(0.06)	(0.01)	(0.06)
$]2,5]$ days × 2015	0.06***	-0.00	0.04	0.00	0.21***	0.00	0.19***
	(0.01)	(0.01)	(0.04)	(0.01)	(0.06)	(0.01)	(0.05)
$]1,4]$ weeks × 2015	0.04***	0.02	0.01	0.01	0.23***	0.01	0.25***
	(0.01)	(0.01)	(0.04)	(0.01)	(0.06)	(0.01)	(0.06)
>4 weeks × 2015	-0.06***	0.01	-0.04	0.01	-0.01	0.04	0.09
	(0.01)	(0.02)	(0.04)	(0.01)	(0.06)	(0.02)	(0.06)
Constant	0.74***	2.19***	2.35***	0.77***	0.30**	0.51***	0.17
	(0.03)	(0.03)	(0.04)	(0.03)	(0.11)	(0.04)	(0.10)
R^2	0.15	0.58	0.52	0.17	0.24	0.23	0.25
Adj. R^2	0.15	0.58	0.52	0.17	0.22	0.23	0.24
Num. obs.	22,860	30,018	4,272	20,012	2,848	20,012	2,848
RMSE	0.22	0.26	0.28	0.17	0.44	0.26	0.43

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Standard errors are in parentheses below estimates. All regressions include controls, for which a comprehensive list can be found in the Appendix. The sample is the same as the sample used to estimate the structural model. Workers with wages below the 1st or above the 99th percentiles are dropped.

4 The basic model

In this section, we develop the framework without training duration. It is indeed straightforward to extend the model and the identification argument to allow for training duration. Moreover, there may exist observed controls, which we condition on throughout.

We assume that workers can be clustered into a finite number K of groups: $k \in \{1,...,K\}$. Let $\pi(k,z,d)$ be the probability mass of the worker type $k \in \{1,...,K\}$, firm's training information policy $z \in \{0,1\}$,[11] and treatment $d \in \{0,1\}$.

We observe three wages $w_t, t = 1,2,3$, per worker. Wage w_1 is the wage before treatment, and w_2, w_3 are two consecutive wages after treatment. We assume that the wage process is Markovian and independent of the instrument given type and treatment. Let $f_t(w_t|k,d)$ denote the density (or probability mass) function for the marginal distribution of wages w_t, with $f_1(w_1|k,d) = f_1(w_1|k)$ independent of d. Let $f_{t|s}(w_t|w_s,k,d)$ denote the density function for the conditional distribution of w_t given w_s (we use $s = t \pm 1$).

Note that wage distributions do not vary with z. The policy variable is therefore an instrument for the treatment. We further assume that pre-treatment wages are independent of both instrument and treatment (z,d) given type k.

This model compares to the Roy model used by Heckman and Vytlacil (2005); Carneiro et al. (2010, 2011). The main difference is that we explicitly model the dependence between model errors via the latent factor k, which is moreover assumed discrete. Specifically, the Roy model consistent with our framework would be similar to

$$\begin{aligned} y &= y(0) + [y(1) - y(0)]\, d \\ d &= 1 \text{ if } \mathrm{E}[y(1) - y(0)|k] \geq c(k,z), \end{aligned}$$

where $y(0), y(1)$ denote post-treatment wages for $d = 0$ or 1, and where $c(k,z)$ is a random training cost given k,z. It is also related to the difference-in-difference model, as we also observe wages before treatment.

4.1 Identification

We make the following assumption on $\pi(k,z,d)$ and $f_2(w_2|k,d)$ and $f_{t|2}(w_t|w_2,k,d)$. Let $\mathcal{W}_2(k,d)$ be the support of $f_2(w_2|k,d)$, and let $\mathcal{W}_2(d) = \bigcap_{k=1}^{K} \mathcal{W}_2(k,d)$ be the common support.

Assumption 1. *For all d,*

a) $\pi(k,0,d) \neq 0$ *for all k;*

b) $[f_{t|2}(w_t|w_2,1,d),...,f_{t|2}(w_t|w_2,K,d)], t = 1,3,$ *are linearly independent systems for all d and for all* w_2 *in* $\mathcal{W}_2(d)$*;*

[11] Binary variable indicating if the worker reports receiving information through any of the following channels: hierarchy, training or HR manager, coworkers, or staff representatives.

c) $\frac{\pi(k,1,d)}{\pi(k,0,d)} \neq \frac{\pi(k',1,d)}{\pi(k',0,d)}$ *for all* k, k'.

Assumption a) means that workers of all types have positive probability of being treated or non treated for at least one instrument value, arbitrarily set equal to 0. Assumption b) means that none of the conditional distributions $f_{t|2}(w_t|w_2,k,d)$, $k=1,...,K$, can be equated with a linear mixture of the other ones. Assumption c) requires different expositions to the instrument for all types.

Assumption 2 (Large support)**.** *For all* d*, distributions* $f_2(w_2|k,d)$*,* $k=1,...,K$*, have identical support:* $\mathcal{W}_2(k,d) \subseteq \mathcal{W}_2(d) = \bigcap_{k=1}^{K} \mathcal{W}_2(k,d)$*.*

The identification proof goes in four steps.

Step 1: Identifying restrictions. Consider first the joint probability of treatment $d_i = d$, instrument $z_i = z$, and wages: w_{i1} (before treatment) and w_{i2}, w_{i3} (after treatment):

$$\begin{aligned} p(z,d,w_1,w_2,w_3) &= \sum_k \pi(k,z,d)\, f_1(w_1|k)\, f_{2|1}(w_2|w_1,k,d)\, f_{3|2}(w_3|w_2,k,d) \\ &= \sum_k \pi(k,z,d)\, f_2(w_2|k,d)\, \frac{f_1(w_1|k)\, f_{2|1}(w_2|w_1,k,d)}{f_2(w_2|k,d)}\, f_{3|2}(w_3|w_2,k,d) \end{aligned}$$

where $f_2(w_2|k,d) = \int f_1(w_1|k)\, f_{2|1}(w_2|w_1,k,d)\, dw_1$ and

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

Without (much) loss of generality, let us assume that wages take N possible discrete values. For continuous wages, we could project the function $(w_1,w_2) \mapsto P(z,d,w_1,w_2,w_3)$ on a functional basis and adapt the identification argument straightforwardly. Then, store these probabilities (given z,d) into a matrix $P(z,d,w_2) = [p(z,d,w_1,w_2,w_3)]_{w_1 \times w_3}$, where the subscript $w_1 \times w_3$ means that we have the values of w_1 indexing rows and those of w_3 indexing columns. Let also $F_1(d,w_2) = \left[f_{1|2}(w_1|w_2,k,d)\right]_{w_1 \times k}$ be the matrix of pre-treatment wage probabilities, with w_1 indexing rows and k indexing columns. Similarly, let $F_2(d,w_2) = \left[f_{3|2}(w_3|w_2,k,d)\right]_{w_3 \times k}$ be the post-treatment matrix. Lastly, let $D(z,d,w_2) = \mathrm{diag}\left[\pi(k,z,d)\, f_2(w_2|k,d)\right]_k$ be the diagonal matrix with $\pi(k,z,d)\, f_2(w_2|k,d)$ in the kth diagonal entry. We then have,

$$P(z,d,w_2) = F_1(d,w_2)\, D(z,d,w_2)\, F_2(d,w_2)^\top.$$

Step 2: Identification given treatment d and first post-treatment wage w_2. We first proceed to show that $F_1(d,w_2), D(z,d,w_2), F_2(d,w_2)$ are identified. Importantly, $F_1(d,w_2)$ and $F_2(d,w_2)$ are independent of z. So, there are two observable matrices, $P(0,d,w_2)$ and $P(1,d,w_2)$, with the same algebraic structure. Under Assumption 1,

$F_1(d, w_2)$ and $F_2(d, w_2)$ are full-column rank, and the matrix $D(0, d, w_2)$ is invertible for all $w_2 \in \bigcap_{k=1}^{K} \mathcal{W}_2(k, d)$ in the common part of the supports of all distributions $f_2(w_2|k, d)$, $k = 1, ..., K$.

To simplify the notations, let us omit the dependence on (d, w_2). Matrix $P(0, d, w_2) :=$ $P(0)$ has rank K and singular value decomposition: $P(0) = U\Lambda V^\top$, where $U^\top U = I_N$, $V^\top V = I_N$ and Λ is a diagonal matrix of dimension N. The number of non-zero diagonal entries in Λ is equal to the number of groups K. Let Λ_1 be the $K \times K$ 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 $P(0) = U_1 \Lambda_1 V_1^\top$.

Note also that

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

Hence,

$$U_2^\top P(0) = U_2^\top F_1 D(0) F_2^\top = 0_{(N-K)\times N},$$

where $F_1(d, w_2) := F_1$ and $F_2(d, w_2) := F_2$. As $D(0)F_2^\top$ is a full row rank K-by-N matrix, it follows that $U_2^\top F_1 = 0_{(N-K)\times K}$. A similar argument implies that $F_2^\top V_2 = 0_{K\times(N-K)}$.

Let $W = \Lambda_1^{-1} U_1^\top F_1$. Then

$$\Lambda_1^{-1} U_1^\top P(0) V_1 = \Lambda_1^{-1} U_1^\top F_1 \times D(0) F_2^\top V_1 = I_K.$$

Hence $W^{-1} = D(0) F_2^\top V_1$.

It also holds that

$$\Lambda_1^{-1} U_1^\top P(1) V_1 = \Lambda_1^{-1} U_1^\top F_1 D(1, d) F_2^\top V_1 = W D(1) D(0)^{-1} W^{-1},$$

where $P(1, d, w_2) := P(1)$. The diagonal entries of $D(1)D(0)^{-1}$ being distinct by assumption, 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 diagonal matrix Δ such that $\widehat{W} = W\Delta =$ $\Lambda_1^{-1} U_1^\top F_1 \Delta$. Then, $\Lambda_1 \widehat{W} = U_1^\top F_1 \Delta$. As $U_2^\top F_1 \Delta = 0_{(N-K)\times K}$, we also have

$$\begin{pmatrix} \Lambda_1 \widehat{W} \\ 0_{(N-K)\times K} \end{pmatrix} = U^\top F_1 \Delta.$$

Hence,

$$U_1 \Lambda_1 \widehat{W} = U \begin{pmatrix} \Lambda_1 \widehat{W} \\ 0_{(N-K)\times K} \end{pmatrix} = U U^\top F_1 \Delta = F_1 \Delta.$$

That is, $F_1 \Delta = U_1 \Lambda_1 \widehat{W}$ is identified. Moreover, since the rows of F_1 sum to one (each column is a probability distribution), then Δ is identified, and therefore F_1 also.

Lastly, $\Delta \widehat{W}^{-1} = W^{-1} = D(0) F_2^\top V_1$. Applying the same argument as in the previous step, we have that

$$\begin{aligned}
W^{-1}V_1^\top &= \Big(D(0)F_2^\top V_1, 0_{K\times(N-K)}\Big)\begin{pmatrix} V_1^\top \\ V_2^\top \end{pmatrix} \\
&= \Big(D(0)F_2^\top V_1,\ D(0)F_2(d)^\top V_2\Big)V^\top \\
&= D(0)F_2^\top VV^\top \\
&= D(0)F_2^\top.
\end{aligned}$$

The rows of F_2 summing to one, it follows that $D(0)$ and F_2 are identified. Hence $D(1)$ is also identified.

Step 3: Common labeling given d. In the previous step, we have estimated

$$D(1,d,w_2)D(0,d,w_2)^{-1} = \operatorname{diag}\left[\frac{\pi(k,1,d)}{\pi(k,0,d)}\right]_k.$$

By assumption 1c, these eigenvalues are all different (and independent of w_2). One can thus relabel groups for each d so that the labelling is consistent for all possible choices of w_2. By Assumption 2, Step 2 can be done for all wages w_2 in the common support, which is also the entire support. Thus, we can sum $D(0,d,w_2)$ and $D(1,d,w_2)$ over w_2 and eliminate $f_2(w_2|k,d)$ (which sums to one on its support). This identifies $\pi(0,k,d)$ and $\pi(1,k,d)$ for all k. Since $f_{1|2}(w_1|w_2,k,d)$ is already identified, then $f_1(w_1|k)\,f_{2|1}(w_2|w_1,k,d)$ is also identified. Summing this function over w_2 then identifies $f_1(w_1|k)$. And the conditional distributions $f_{t|t-1}(w_t|w_{t-1},k,d), t=2,3$ finally follow.

Step 4: Common labeling across treatments. It remains to align the groupings across treatments. This is easily done by remarking that $f_1(w_1|k)$ is independent of d and therefore can be used to make sure that the same groups have identical labels across treatments.

5 Estimation

Wages are assumed log-normal given match type and training (no training, and training by duration). That is,

$$f_1(w_1|k) \equiv \frac{1}{w_1}\frac{1}{\sigma_1(k)}\varphi\left(\frac{\ln w_1 - \mu_1(k)}{\sigma_1(k)}\right), \tag{2}$$

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

$$f_{2|1}(w_2|w_1, k, d) = \frac{1}{w_2} \frac{1}{\sigma_2(k,d)} \varphi\left(\frac{\ln w_2 - \mu_2(k,d) - \rho_2(k,d)\,[\ln w_1 - \mu_1(k)]}{\sigma_2(k,d)} \right),$$
$$f_{3|2}(w_3|w_2, k, d) = \frac{1}{w_3} \frac{1}{\sigma_3(k,d)} \varphi\left(\frac{\ln w_3 - \mu_3(k,d) - \rho_3(k,d)\,[\ln w_2 - \mu_2(k,d)]}{\sigma_3(k,d)} \right).$$

And we use discrete probabilities for $\pi(k, z, d)$.

Let β gather all parameters. The data for each individual is $x_i = (w_{i1}, w_{i2}, z_i, d_i)$. The individual complete likelihood (i.e. of data and unknown type k) is then

$$\ell(x_i, k|\beta) \equiv \ell_i(k|\beta) = \pi(k, z_i, d_i)\, f_1(w_{i1}|k)\, f_{2|1}(w_{2i}|w_{1i}, k, d_i)\, f_{3|2}(w_{3i}|w_{2i}, k, d_i). \tag{3}$$

The individual likelihood is $\ell_i(\beta) = \sum_k \ell_i(k|\beta)$. The sample likelihood is $L(\beta) = \prod_i \ell_i(\beta)$.

5.1 The EM algorithm

We use the EM-algorithm to maximize the sample likelihood. The algorithm iterates the following two steps.

E-step. For a given value $\beta^{(m)}$ of the parameter, the posterior probability of worker i to be of type k given data (also called *responsibility*) is

$$p_i(k|\beta^{(m)}) \equiv \frac{\ell_i(k|\beta^{(m)})}{\sum_k \ell_i(k|\beta^{(m)})}. \tag{4}$$

M-step. The M-step maximizes the expected complete log-likelihood (given data)

$$\sum_i \sum_k p_i(k|\beta^{(m)}) \ln \ell_i(k|\beta).$$

We update parameters sequentially as follows.

1. Update pre-treatment wage distribution parameters as

$$\mu_1^{(m+1)}(k) = \frac{\sum_i p_i(k|\beta^{(m)}) \ln w_{i1}}{\sum_i p_i(k|\beta^{(m)})},$$
$$\sigma_1^{(m+1)}(k) = \sqrt{\frac{\sum_i p_i(k|\beta^{(m)})[\ln w_{i1} - \mu_1^{(m+1)}(k)]^2}{\sum_i p_i(k|\beta^{(m)})}}.$$

2. Then, separately for each k and d, regress $\ln w_{i2}$ on $\ln w_{i1} - \mu_1^{(m+1)}(k)$, weighting each observation by $p_i(k|\beta^{(m)})\mathbf{1}\{d_i = d\}$. The regression's intercept is $\mu_2^{(m+1)}(k, d)$ and the RMSE is $\sigma_2^{(m+1)}(k, d)$.

3. Then, again separately for each k and d, regress $\ln w_{i3}$ on $\ln w_{i2} - \mu_2^{(m+1)}(k,d)$, weighting each observation by $p_i(k|\beta^{(m)})\mathbf{1}\{d_i = d\}$. The regression's intercept is $\mu_3^{(m+1)}(k,d)$ and the RMSE is $\sigma_3^{(m+1)}(k,d)$.

4. Finally update state probabilities as

$$\pi^{(m+1)}(k,z,d) = \frac{1}{N} \sum_{i:z_i=z,d_i=d} p_i(k|\beta^{(m)}).$$

Standard errors. Let

$$S_i(\beta) = \frac{\partial \ln \sum_k \ell_i(k|\beta)}{\partial \beta}$$

be the gradient of the individual log-likelihood. We estimate the asymptotic variance of a MLE $\widehat{\beta}$, as

$$\widehat{V} = \left[\sum_i S_i(\widehat{\beta})\, S_i(\widehat{\beta})^\top\right]^{-1}.$$

Jamshidian and Jennrich (2000) discuss various other ways to calculate standard errors, but this general, standard one is always amongst the most precise.

In order to calculate the gradients S_i one will use numerical derivatives, specifically a first-order Richardson extrapolation of the central difference:

$$f'(x) \simeq \frac{f(x-2h) - 8f(x-h) + 8f(x+h)}{12h}$$

where $h_j = \eta \max(|\beta_j|, 1)$ for the jth coordinate and with $h = 10^{-7}$. The simulations in Jamshidian and Jennrich (2000) show that it dominates the usual forward difference.

Then, any (continuous) parameter transformation $b(\beta)$ can be estimated as $b(\widehat{\beta})$ and its asymptotic variance can be calculated using the delta-method (if continuously differentiable):

$$\widehat{\mathrm{Var}}\, \widehat{b} = \frac{\partial b(\widehat{\beta})}{\partial \beta} \widehat{V} \left[\frac{\partial b(\widehat{\beta})}{\partial \beta}\right]^\top.$$

6 Treatment effects and usual estimators

6.1 Same training duration

We first consider the simpler case where there the treatment is binary. We shall incorporate back training duration in the next subsection.

Let $y(0)$ denote the logged wage in period $t = 2,3$ for untrained workers ($d = 0$), and let $y(1)$ the wage for trained workers ($d = 1$). Let also $y = d\,y(1) + (1-d)\,y(0)$. Let $z = 0$ (no information) or $z = 1$ (information on training made available by employer) be the instrument.

ATE. The conditional Average Treatment Effect is

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

The unconditional ATE is

$$ATE = \sum_k \pi(k)\, ATE(k),$$

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

ATT. The unconditional Average Treatment on the Treated is

$$\begin{aligned} ATT &= \mathrm{E}[y(1)|d=1] - \mathrm{E}[y(0)|d=1] \\ &= \sum_k \pi(k|d=1)\, ATE(k) \end{aligned}$$

with $\pi(k|d) = \sum_z \pi(k,z|d)$ for

$$\pi(k,z|d) = \frac{\pi(k,z,d)}{\sum_{k,z} \pi(k,z,d)}, \ d = 0,1.$$

OLS. The OLS estimator of the regression of y on d is

$$\begin{aligned} b_{OLS} &= \frac{\mathrm{Cov}(y,d)}{\mathrm{Var}(d)} = \mathrm{E}[y(1)|d=1] - \mathrm{E}[y(0)|d=0] \\ &= \sum_{k,z} \pi(k,z|d=1)\, \mu(k,1) - \sum_{k,z} \pi(k,z|d=0)\, \mu(k,0) \\ &= ATT + \sum_k [\pi(k|d=1) - \pi(k|d=0)]\, \mu(k,0). \end{aligned}$$

The DiD estimate is the OLS estimator for $y(1), y(0)$ defined as wage increases between before and after the treatment's application.

IV and LATE. Finally, the IV estimator of the regression of y on d using z as instrument is

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

The numerator can be calculated as

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

making use of

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

The extra term $\sum_k \left[\pi(k|z=1)-\pi(k|z=0)\right]\mu(k,0)$ makes IV here differ from LATE because unobserved heterogeneity correlates with the instrument. Lastly,

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

The IV estimator is thus a weighted mean of conditional ATEs, $\mu(k,1)-\mu(k,0)$. This average is informative if the weights are uniformly positive or negative (monotonicity; see Imbens and Angrist, 1994). 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 assume monotonicity in the treatment probability.

Hence, as with the Roy model studied by Heckman and Vytlacil (2005), standard estimates fail to deliver easily interpretable information on the distribution of treatment effects.

6.2 Multivalued treatment

ATE. With variable training duration, the treatment is no longer binary and the outcome is $y(d)$ where $d=0$ indicates no training and $d=1,2,...>0$ indicates different training durations. We can then redefine the conditional ATE as

$$ATE(k,d)=\mathrm{E}(y(d)|k)-\mathrm{E}(y(0)|k)=\mu(k,d)-\mu(k,0),\ d>0.$$

It is also possible to average over training duration as

$$\begin{aligned}
ATE(k) &= \mathrm{E}(y(d)|k,d>0)-\mathrm{E}(y(0)|k) \\
&= \frac{\sum_z \sum_{d>0}\pi(k,z,d)\mu(k,d)}{\sum_z \sum_{d>0}\pi(k,z,d)}-\mu(k,0).
\end{aligned}$$

The definitions of unconditional ATE and ATT is as before.

IV. The IV estimator straightforwardly generalizes. We can still calculate

$$b_{IV} = \frac{\mathrm{Cov}(y,z)}{\mathrm{Cov}(D,z)} = \frac{\mathrm{E}(y|z=1)-\mathrm{E}(y|z=0)}{\mathrm{E}(D|z=1)-\mathrm{E}(D|z=0)}$$

with $D=\mathbf{1}\{d>0\}$. First,

$$\begin{aligned}
&\mathrm{E}(y|z=1)-\mathrm{E}(y|z=0)\\
&\qquad=\sum_k\sum_d \pi(k,d|z=1)\,\mu(k,d)-\sum_k\sum_d \pi(k,d|z=0)\,\mu(k,d)\\
&\qquad=\sum_k\sum_{d>0}\left[\pi(k,d|z=1)-\pi(k,d|z=0)\right]\left[\mu(k,d)-\mu(k,0)\right]\\
&\qquad\qquad+\sum_k\left[\pi(k|z=1)-\pi(k|z=0)\right]\mu(k,0)
\end{aligned}$$

using

$$\pi(k,d=0|z)=\pi(k|z)-\sum_{d>0}\pi(k,d|z).$$

And we have

$$\mathrm{E}(D|z=1)-\mathrm{E}(D|z=0)=\sum_k\sum_{d>0}\left[\pi(k,d|z=1)-\pi(k,d|z=0)\right].$$

The IV estimator continues to be a sensible aggregation of ATEs is the training probabilities are monotonous in the instrument z for all k and d.

7 Results

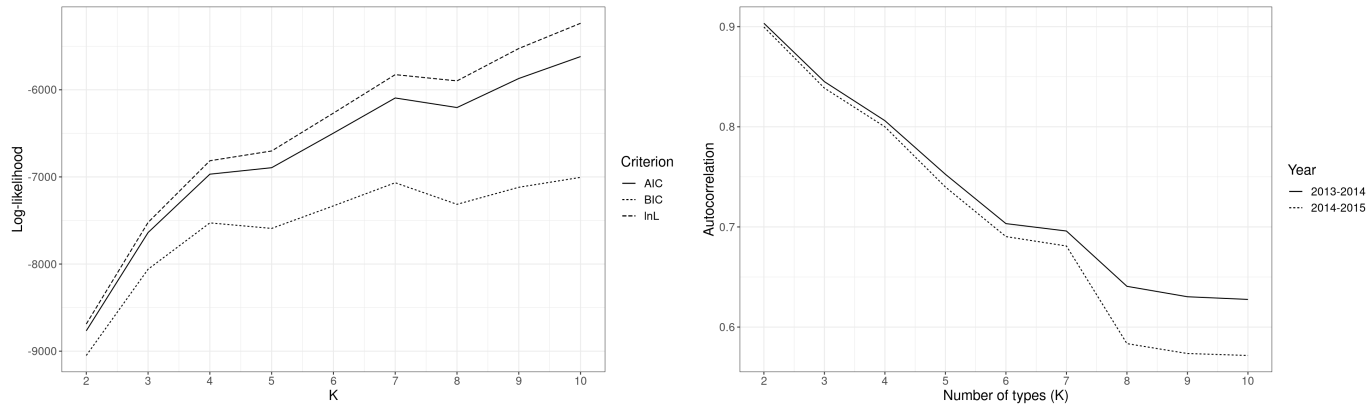
This section presents the results from estimating our model using the EM-algorithm. We first discuss how the number of types (K) affects our results and how we choose this key parameter. We then present estimates of other important parameters, and some treatment effects derived from these estimates. We test some of the key assumptions. Finally, we reintroduce some of the observed variables we have on the workers and their firms, and we analyse how firm and individual characteristics differ across types.

7.1 The number of types (K)

Our estimation strategy requires the econometrician to choose the number of types, K. Understanding what a "type" is in our model and how the number of types affects the estimation is therefore crucial. Individuals draw their wage from distributions conditional on their type in 2013, and both their type and training status in 2014 and 2015. The

Figure 2: Criteria to help in choosing the number of types

(a) (Penalized) log-likelihood (b) Residual autocorrelations



Notes: (a) The AIC and BIC are criteria that penalise the model for having more parameters. If M is the number of parameters, N the number of observations, and L the likelihood, AIC $= 2\ln L - \ln M$, and BIC $= \ln L - \ln(N)M$. (b) For the autocorrelations, rho (ρ) is the autoregressive parameter estimated within the EM algorithm (a parameter of the model), while autoCorr is the calcualted autocorrelation of wage residuals after the estimation algorithm converges.

algorithm uses three factors to classify individuals; the mean wage across the three years, the variance of wages within each type, and the effect of training on wages. More groups allow more flexibility, though at the expense of the interpretability of the (increasing number of) parameters. For example, allocating each individual to their own group, would be a "perfect" classification, but of no analytical use.

Figures 2 and 3 present some criteria we use to choose the number of types for the remainder of our analysis. In figure 2(a), the different lines show how the total likelihood ($\ln L$) and penalised-likelihoods evolve with K. We are looking for "elbows"; points where the gain in likelihood for an additional type is noticeably less than it was previously. There are clear elbows at four types and seven types. We are also after elbows on figure 2(b), though now we want to find when increasing the number of types no longer gives a strong reduction in autocorrelation. The first elbow is at six types and there is another at eight.

Figure 3 shows how the assignment of individuals to types evolves as we increase the number of types. The size of the node is proportional to the number of individuals assigned to that group, while their vertical position is the estimated mean wage in 2013 of that type, $\mu_1(k)$. Our estimation algorithm does not assign individuals to groups, so we derive these assignments as the maximal posterior probability of a given individual across all possible types.[12] Given these type assignments for a given number of types, we can track which individuals are assigned to which types as we increase K. This exercise is visualised in figure 3. We want our grouping to be stable and well-defined, so we do not want our chosen number of types to include "phantom" groups that dissappear

[12]The EM algorithm produces posterior probabilities for each individual and each type, $p_i(k)$. We assign an individual i to type k, if $p_i(k) \equiv \max_{k'} p_i(k')$.

as we increase K, nor do we want too many "similar" groups.[13] Finally, if groups are split between K and $K+1$ and remain split for $K+2$ etc., choosing $K+1$ is preferred. Promising choices are four types which has a stable bottom and middle group (though the other groups are quickly split), and seven types, which seems stable everywhere.

We choose to use both four and seven for the remainder of our analysis, as these best meet our criteria. In addition, analysing the estimates of our model using different numbers of types will help us understand the sensitivity of our results to the choice of this parameter.

7.2 Parameter estimates and treatment effects

A key feature of our model is the conditional wage distributions that individual wages are drawn from. The parameters of these conditional distributions are some of the key parameters that we estimate, and so we directly estimate these distributions under assumptions of normality. Figure 4 displays the estimated distribution functions of these conditional distributions for different years (columns), training durations (rows), and types (colours). Note that the left hand column is necessarily identical in all rows as we assume that all individuals draw from the same distribution in 2013, regardless of training in 2014. We will discuss this assumption in detail later.

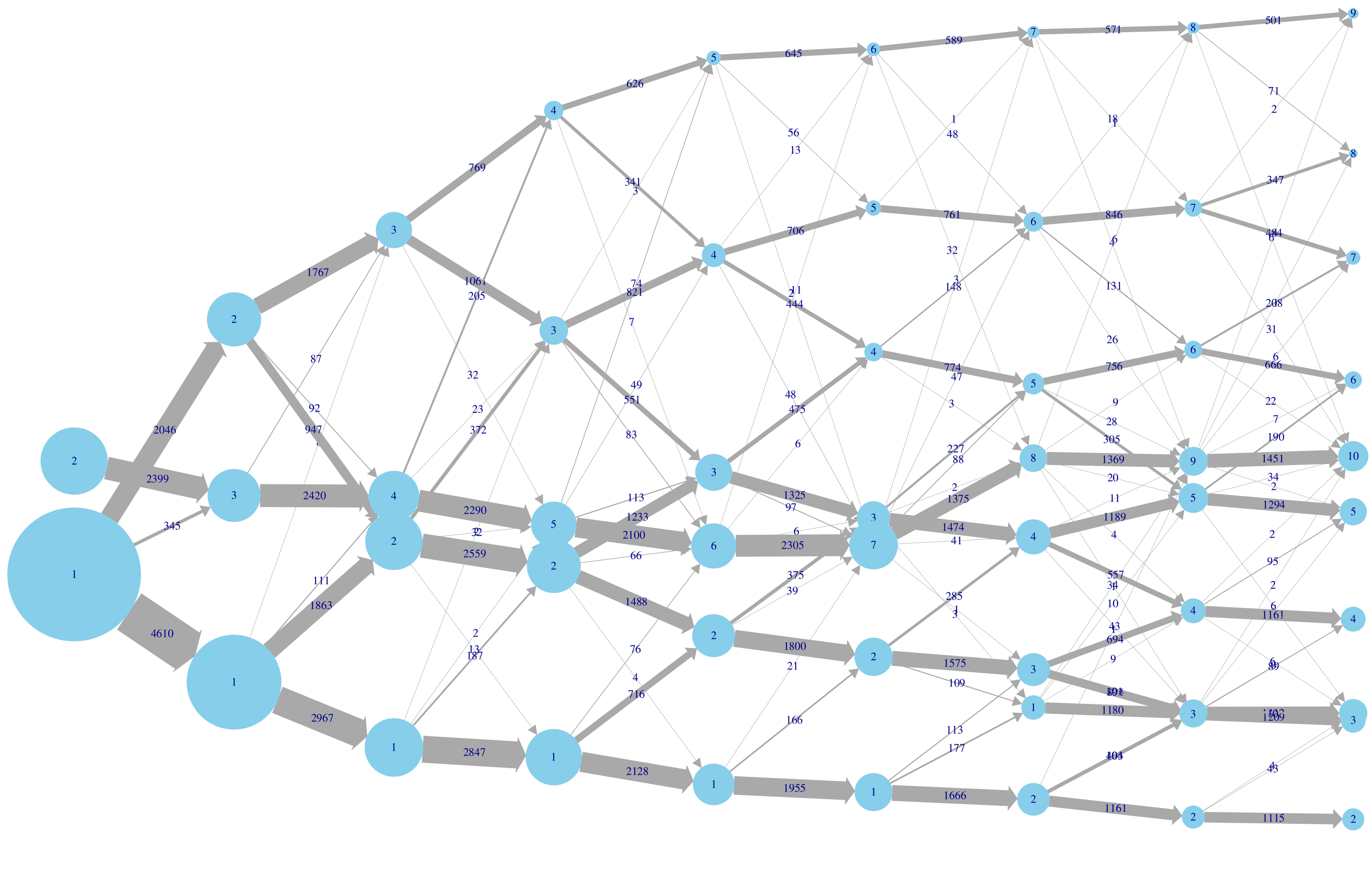
Some of the key features of these distributions, and the mechanism behind the type classification can be seen in figure 4. The types differ mainly by group mean and have similar variances, though at both four and seven types there is (at least) one group with a much higher variance. The within group variation for different years and training durations is difficult to see clearly, but the location and dispersion of the distributions seem relatively stable across rows and down columns. The bottom row for four types is an exception, with type one and type three variances clearly increasing, while the variance of the type four distribution falls. [need to investigate this further perhaps?]

Alongside the location and dispersion parameters of the conditional wage distribution, we also estimate the proportion of individuals in the sample who draw their wage from a given distribution. We called this value $\pi(k, z, d)$ in the model in section 4. Figure 5 displays the estimates for these parameters with four types in 5(a), and seven types in 5(b,c). The picture is generally similar across types, whether there are four or seven. We see a decrease in probability as the duration of training increases, though this is less stark as the group mean wage increases.[14]

Figure 6 demonstrates that our instrument, the receipt of information on training by a worker, is an important predictor of training. For all but one type, an individual who

[13] Groups can be similar along three dimensions: mean wage, variance of wages, and effect of training. Only the first of these dimensions is apparent in figure 3.

[14] Mean wage is increasing in type to $k = K - 1$, with the "top" type ($k = K$) having a mean wage close to that in the full sample, but the highest variance.

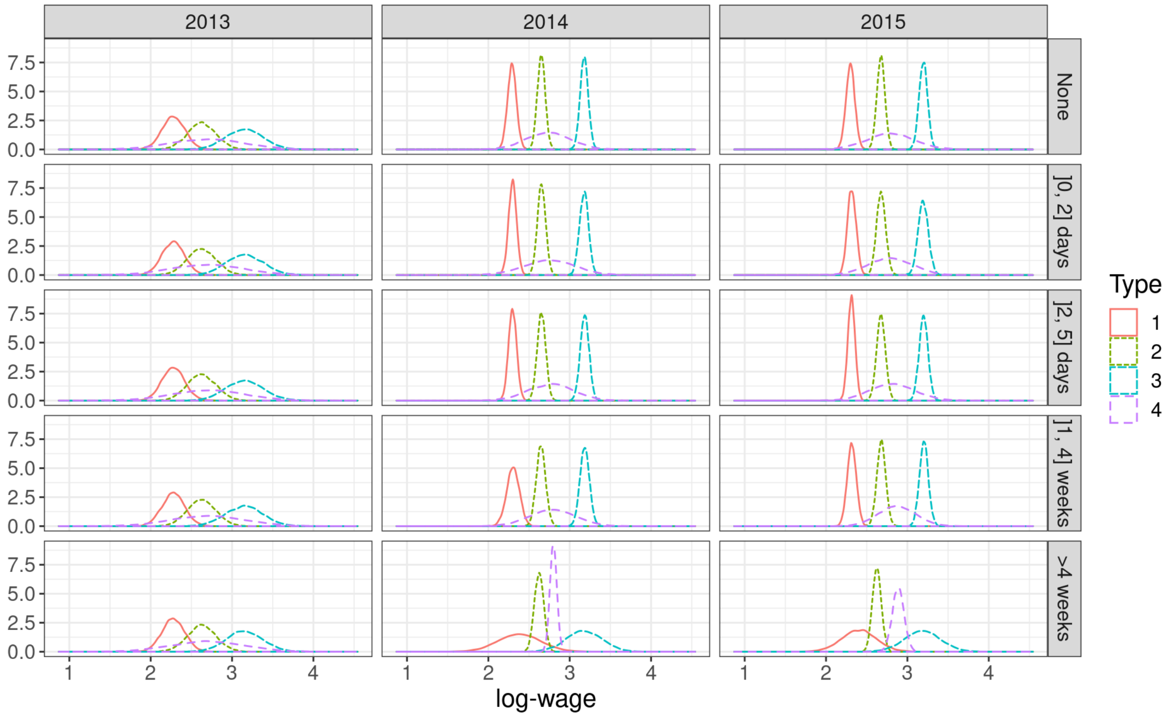


Notes: The size of nodes corresponds to the number of individuals allocated to that type. The nodes are organised horizontally by number of types (K), and vertically by mean wage of that type ($\mu_1(k,0)$).

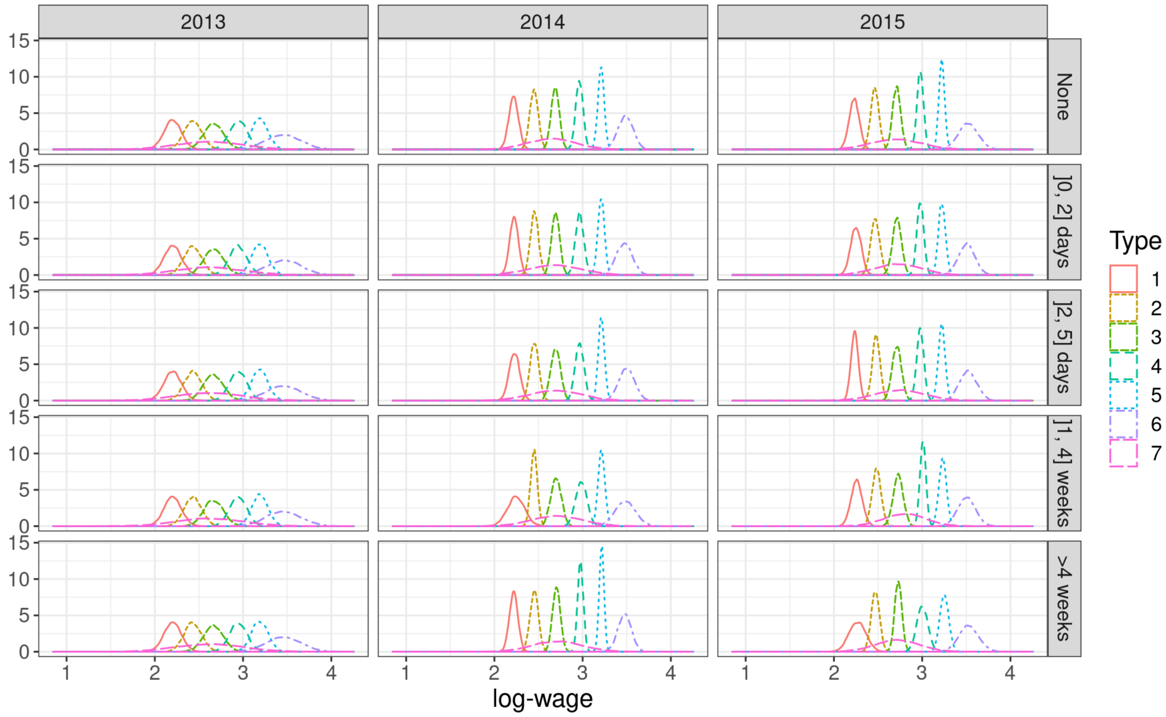
Figure 3: Group branching

Figure 4: Estimated PDFs of wages

(a) $K = 4$



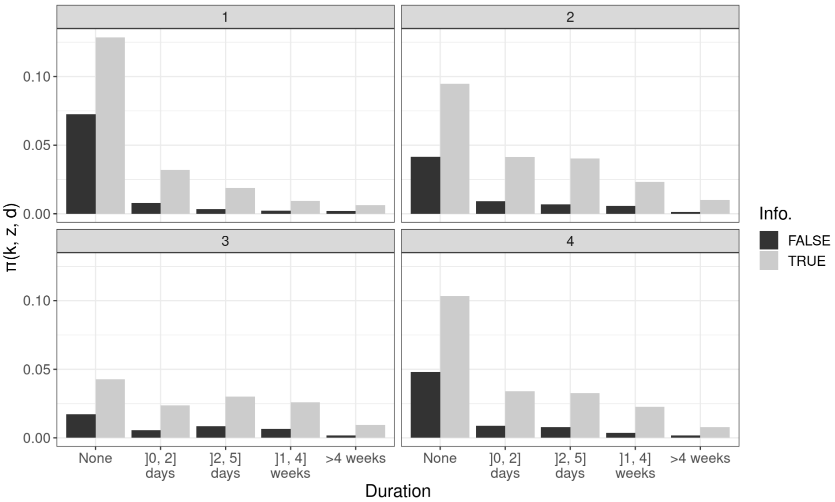
(b) $K = 7$



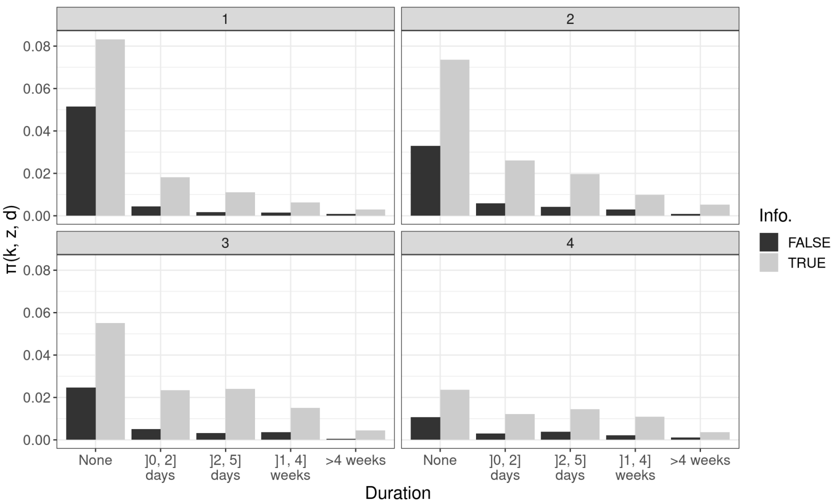
Notes: These are the estimated conditional wage distributions by year (columns), training duration (rows), and type (colours). These distributions were plotted by drawing 10,000 wages from normal distributions with the estimated parameters from the model, with the aid of the R function `rnorm()`.

Figure 5: Distribution of types, training and information

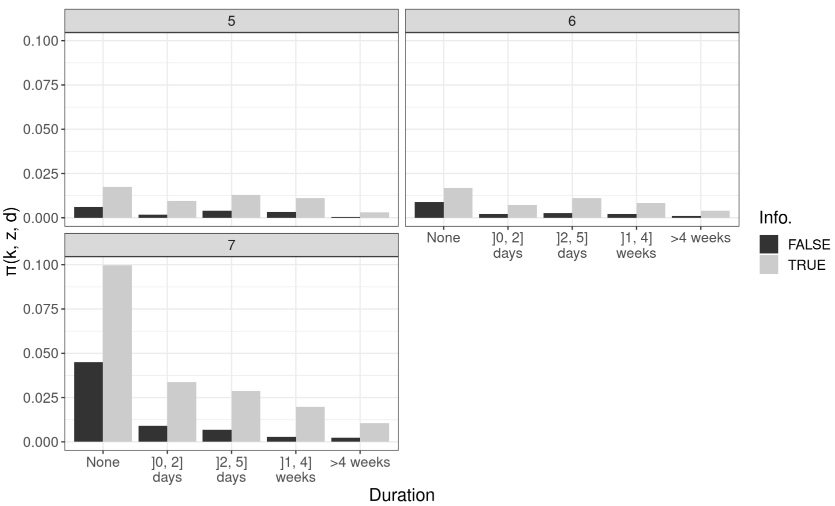
(a) $K = 4$



(b) $K = 7, k = 1, ..., 4$



(c) $K = 7, k = 5, 6, 7$



Notes: The bars represent the proportion of individuals in the sample estimated to belong to a given type, k, by their information (z) and training duration (d) status. Each cell corresponds to a different type, with training duration along the x-axis. The grey bar represents individuals who did not receive information on training, and the yellow those who did. Summing $\pi(k,z,d)$ over k, z, and d will give one.

received information on training is also more likely to receive training of any duration. Our strategy does not require the effect of the instrument on treatment to be monotone, unlike a traditional instrumental variable estimation strategy.

In addition to the parameters that are estimated directly by our implementation of the EM algorithm, we can also derive analogues of some commonly estimated *treatment effects* from our estimated parameters. Formal derivations of these treatment effects are presented in section 6. We calculate the estimated *average treatment effect conditional on type and training duration, $ATE(k, d)$*, with four types in figure 7(a) and seven types split between figures 7(b) and 7(c). These ATEs are differences between log-mean wages, and so can be interpreted as percentage changes in wages. The plots in figure 7 are arranged with each cell corresponding to a type, each column a training duration, and the different bars representing 2014 (grey) and 2015 (yellow). There is no column for "no training" as this is the reference category against which the effects are calculated. Also on the plots are both the proportions within each type-training group, $p(k, d)$, and the mean wage in 2013 for that type, $\mu_1(k, d)$.

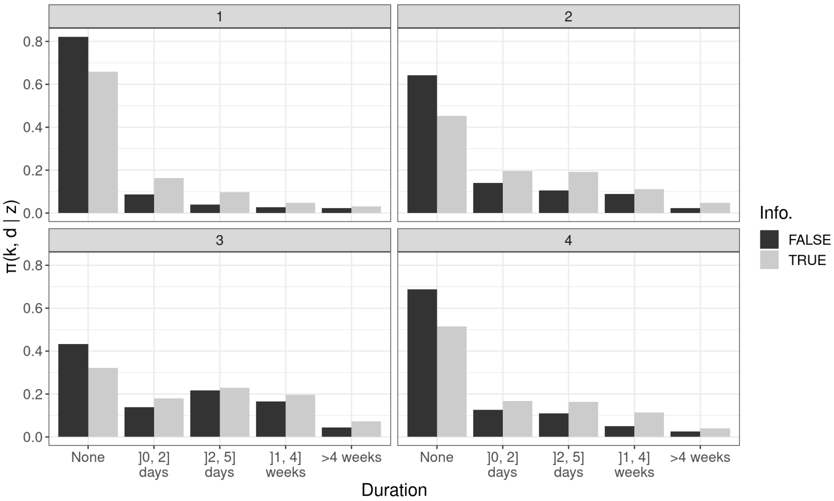
Focusing first on the model with four types in figure 7(a), a few type-training combinations stand out. The benefits from most durations of training are negligible for type one workers, except training of longer than four weeks. Workers who received training for this length of time in 2014 received on average around a ten percent higher wage in both 2014 and 2015. Curiously, type two workers who trained for more than four weeks in 2014 received a *lower* wage than their untrained counterparts in both 2014 and 2015. Other durations of training had little effect on wages for type two and three workers. Finally, type 4 workers benefit from all durations of training.[15] For the two shorter durations, the wage increase over the untrained is approximately two percent for both years. Type four workers who received over one week of training received wages over five percent higher than their untrained counterparts in 2014, rising to over eight percent higher in 2015. The wage benefits of training are also increasing in training duration for this group.

Moving to seven types changes the picture somewhat. There is no longer a type that suffers from the longest duration of training. The effects for types one and two appear to not follow much of a trend, either by year or duration. Types three, four, and five benefit slightly from longer durations of training, with wages around two percent higher for those training for more than one week relative to the untrained. However, the benefits of shorter durations are negligible for these types. The effects of training for type six workers vary significantly, include some apparent negative effects of training on wages. Type seven displays positive effects of training on wages for all durations, and these benefits are increasing for durations up to four weeks at which point they level off in 2014, and fall in 2015. As with the model with four types, we see the most consisitent benefits for the

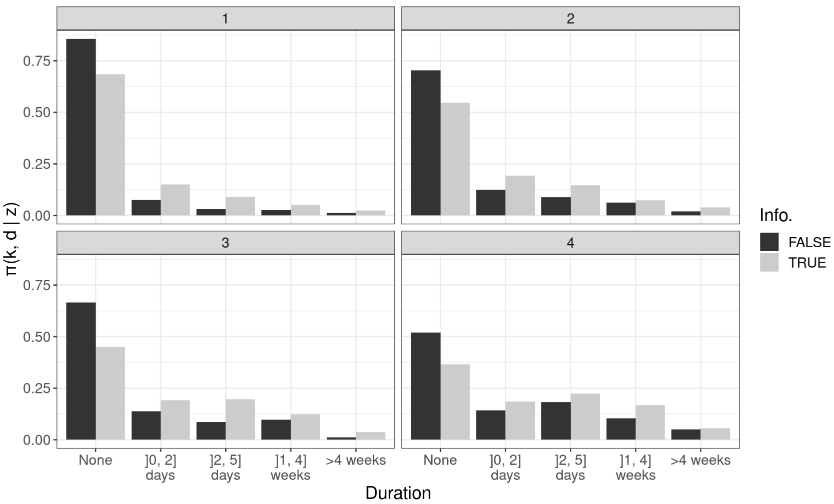
[15]Recall that type 4 is not the group with the highest mean wage, but rather the highest within-group variance, and a mean wage around the population mean.

Figure 6: Distribution of types and training by information status

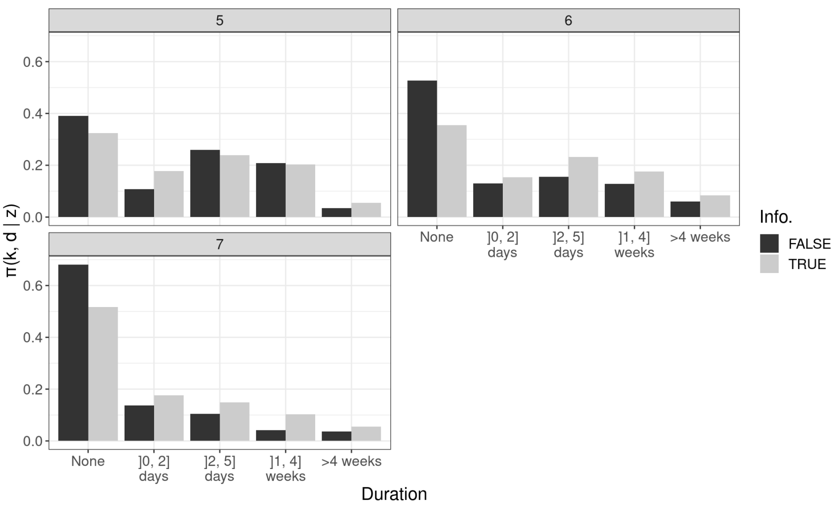
(a) $K = 4$



(b) $K = 7, k = 1, ..., 4$



(c) $K = 7, k = 5, 6, 7$



Notes: These plots are very similar to those in figure 5, with one key difference. The bars now represent proportions of individuals of an estimated type and training, *conditional on their information status.*

Table 5: Model estimates of unconditional treatment effects

	ATE		*ATT*		*IV*		*LATE*	
K	4	7	4	7	4	7	4	7
2014	0.014	0.015	0.031	0.020	0.238	0.231	0.011	0.014
2015	0.017	0.018	0.038	0.025	0.246	0.238	0.016	0.020

Notes: These treatment effects are obtained from the estimated parameters of the model as described in section 6.

group with a high variance, and mean close to that of the whole sample.

7.3 Estimating traditional treatment effects

7.4 Testing the similarity of wage distributions of trained and untrained workers

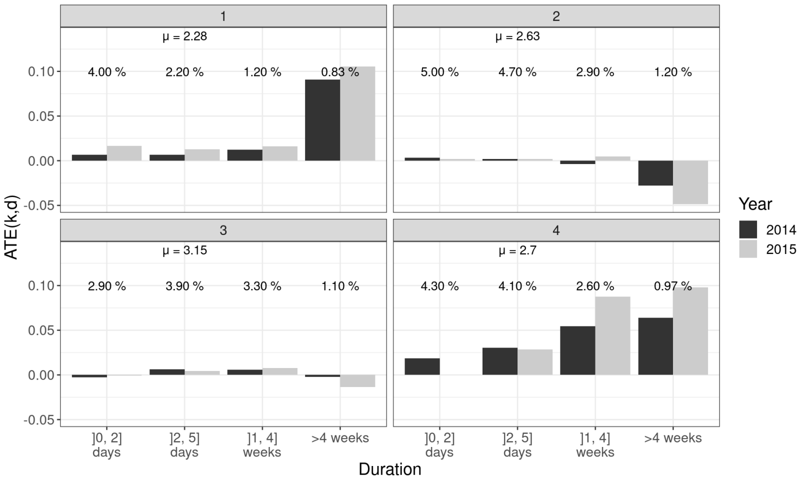
An assumption that we wrestled with often was the equality of wage distributions for trained and untrained workers *before* our training period, i.e. in 2013. This assumption initially made sense given what is generally known about training; if there is any wage increase associated with training it will be due to the effect of training on worker productivity, and so the effects will necessarily occur post-training. However, our reduced-form results, especially the differences-in-differences estimates, suggested that workers who were trained in 2014 were already being paid more in 2013. If this is due to selection on unobserved variables, then that is exactly what our model is designed to account for. Conversely, if this is due to pre-effect of training on wages, perhaps due to the timing of wage increases and training, or training being *written in* to contracts and hence pre-compensated, then this is an important effect of training that we should allow our model to capture. We experimented with a model (for which the results are available on request), that allowed pre-training wage distributions to vary with training. Unfortunately this model posed a more serious problem; without any trained and untrained individuals drawing from the same distributions, the model said nothing about how types for trained individuals corresponded to types for untrained individuals. This meant we could now *choose* whether to see a pre-training effect on wages, and if we and if we do not want to assume the equality of wages conditional on type in 2013, we have to find another way of matching types. We decided to stick with our original assumption, and hence the original model.

We assume all individuals of the same type draw their wages from the same distribution in 2013. Then, if there is an effect of training, untrained and trained individuals *of the same type*, should draw their wages from different distributions in 2014 and 2015.

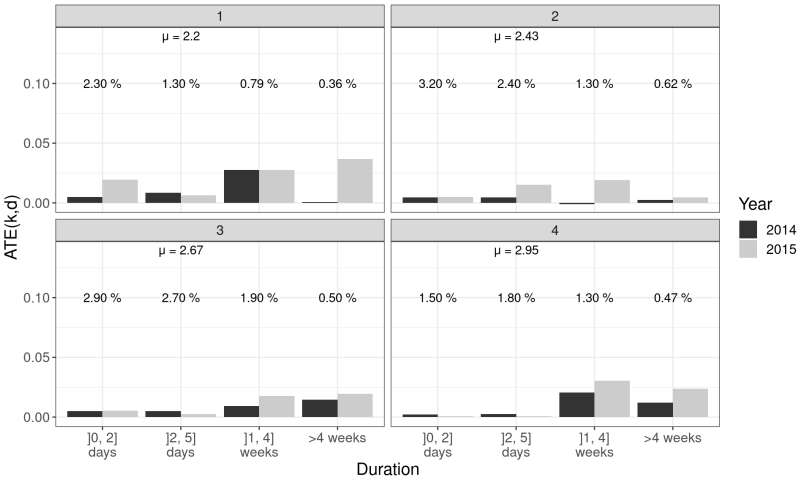
We can compare distributions by examining quantile-quantile (Q-Q) plots. Their name

Figure 7: Average treatment effects by type and training duration, $K = 4$

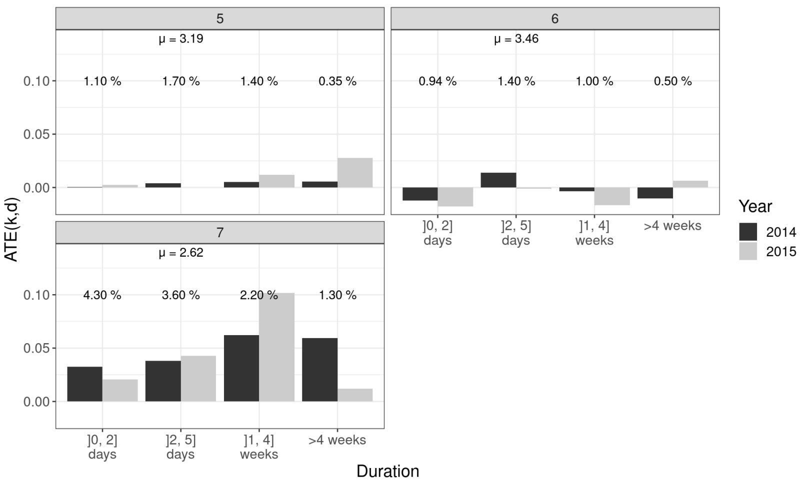
(a) $K = 4$



(b) $K = 7, k = 1, ..., 4$



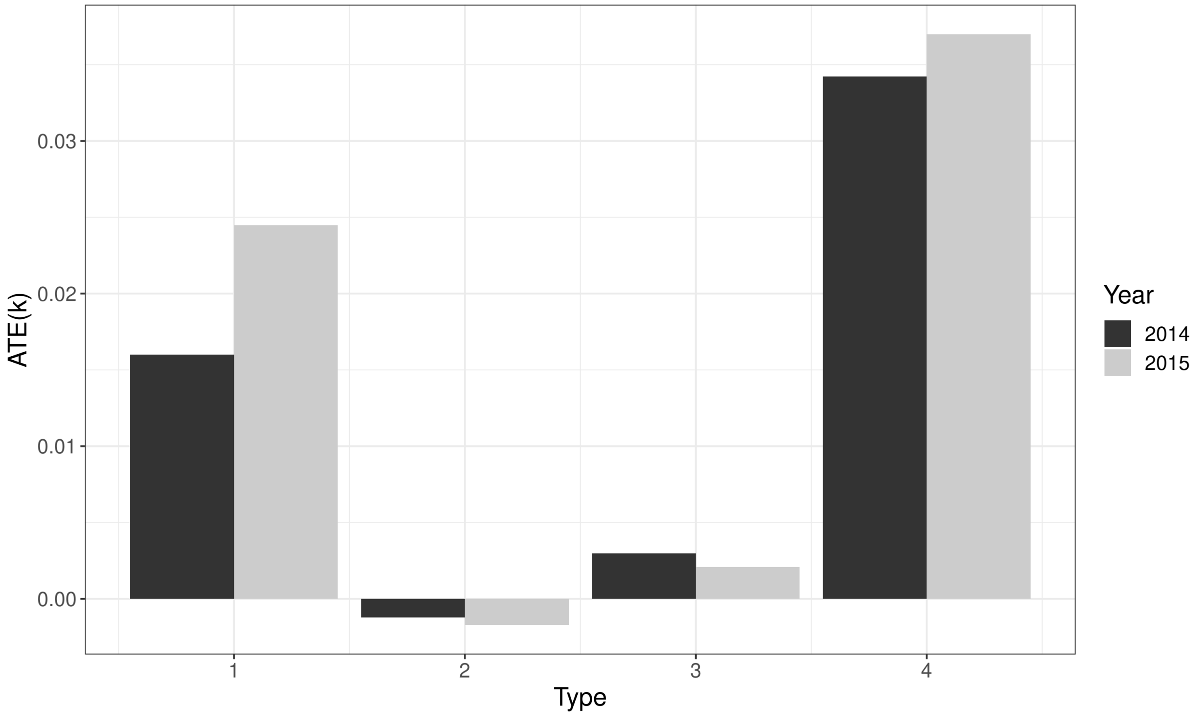
(c) $K = 7, k = 4, ..., 7$



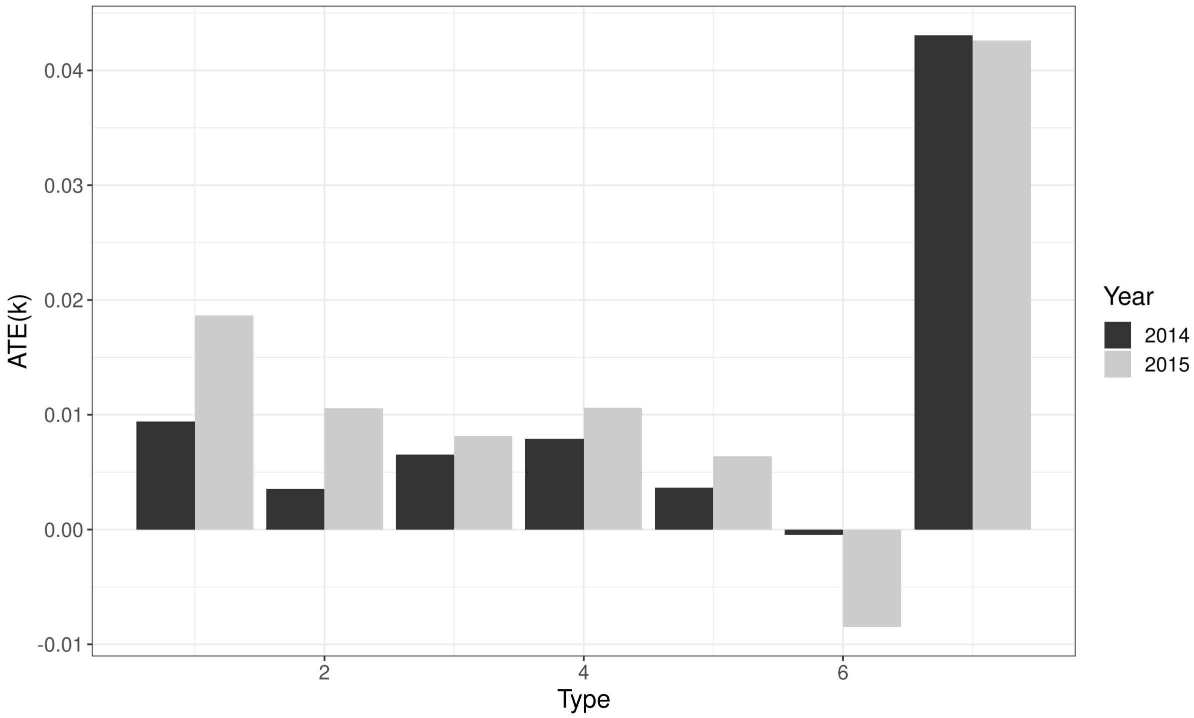
Notes: The average treatment effects are calculated as $\mu_t(k,d) - \mu_t(k,0)$, where $t = 2014, 2015$ and $d = 1,2,3,4$. Also displayed on the plot are the shares of individuals in each category. So looking at the left-most column in the bottom-right cell, the probability of an individual chosen at random in the sample being of type 4, and receiving two days or less of training is 4.30%. Each cell also displays the 2013 *unconditional* mean wage for that type, $\mu_1(k)$.

Figure 8: Estimated average treatment effects by type

(a) $K = 4$

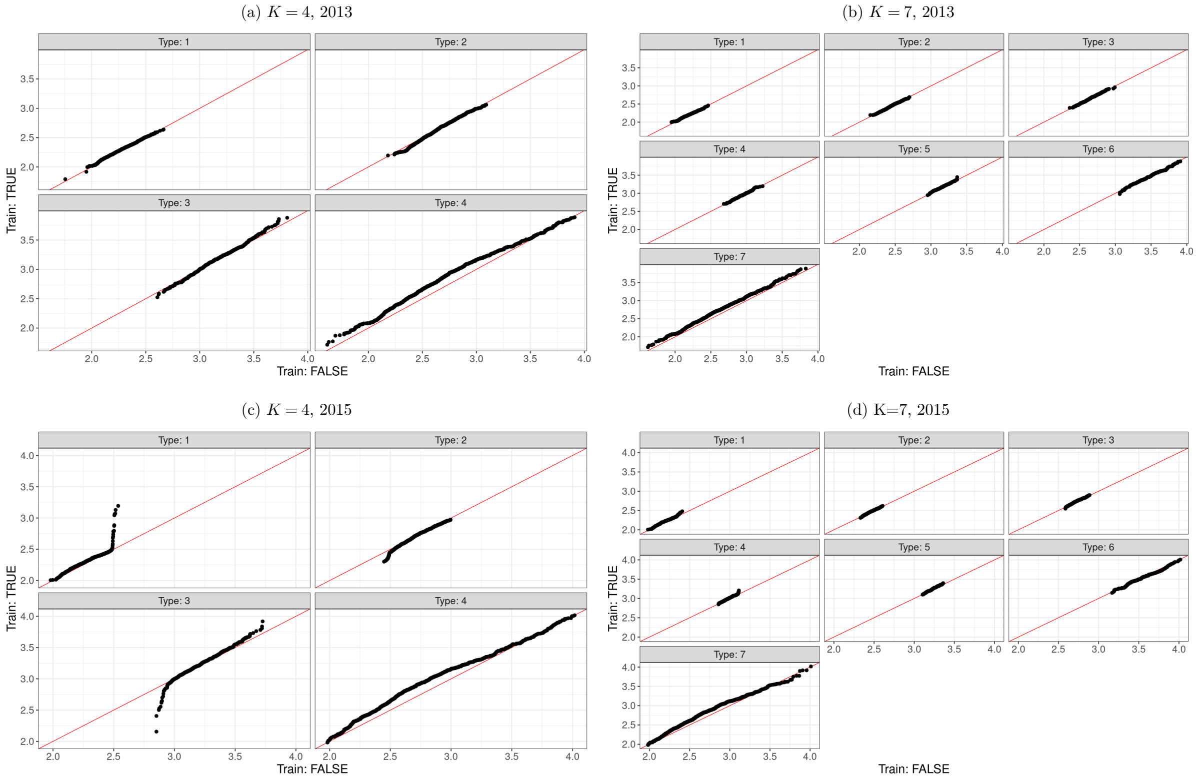


(b) $K = 7$



Notes: The average treatment effects are calculated as described in section 6, and are conditional on type.

Figure 9: Q-Q plots for trained vs untrained workers' wages



Notes: The red line is the 45-degree line through the origin. The more closely the plotted points follow the red line, the closer are the wage distributions for trained and untrained.

is self-explanatory; Q-Q plots plot the quantiles of one (empirical) distribution on the x-axis against the quantiles of another on the y-axis. The more similar are the distributions the closer the points will be to lying on the line $y = x$. Figure 9 does exactly this for the wages of untrained (x-axis) and trained (y-axis) workers in 2013 and 2015, for four and seven types. The 2013 points are perhaps slightly closer to the red line ($y = x$), especially with four types. However, there does not appear to be a clear difference between 2013 and 2015 for most types, so if this exercise does support our assumption regarding pre-training wages, it does not provide much support for an effect of training on post-training wages.

7.5 Individual and firm characteristics by type

The results we have presented in this section have mostly been conditional on worker *types* though we have not discussed in detail what types represent. Worker types capture (unobserved) worker heterogeneity that affect wages independently of training, and variation in the probability of, and responses to, training. We place "unobserved" in parentheses because, although the technique we use is typically employed to capture unobserved heterogeneity, there is no reason why it cannot also capture heterogeneity that we can measure otherwise. By omitting controls from our estimation these types capture *all* the important variation, both observed *and* unobserved. We do this to keep the number of parameters in the model reasonable, so they remain interpretable and precisely estimated. However, *ex post* we can compare the characteristics we do observe across assigned types. These characteristics by type are displayed for the model with seven types in table 6.

The top two rows of table 6 correspond to empirical mean and standard deviation of 2013 wages for workers assigned to that type. Note this may differ slightly from the estimated group means and standard deviations as they are weighted means using the posterior probabilities obtained in the EM algorithm—they are also in EUR, while we worked with the natural logarithm of wages in the model. We can see that both the mean and the standard deviation of wages are increasing in type, up to type six. Type seven is somewhat of an anomaly, and hence is separated in the table. The mean and variance of type seven wages are close to the full sample mean and variance.

Ignoring type seven for now, if we split the individual characteristics in the table into "positive" and "negative" characteristics, the positive ones are increasing with type, while the negative are decreasing. For example, fifty-seven percent of type one workers have a highest qualification less than the *baccalauréat*, while this is true for only nine percent of type six workers—eight-two percent of whom have a qualification corresponding to at least two years of higher education. A similar pattern emerges for firms. The positive characteristics (e.g. incentives), are increasing in type, while more "negative ones" (e.g. percent with short-term contract at the firm) are decreasing. Higher type workers are also more likely to work in bigger firms, measured by number of employees. The type-sector matching reflects the U-shape described in the literature on job polarisation (Autor et al., 2003), with both low and high types more likely to work in services than middle types—though manufacturing workers are a minority in all types.

Returning to type seven, there are not many characteristics that stand out other than the variance of wages. Like the mean of wages, other characteristics for this type fit somewhere in the middle of the other six types, usually somewhere between types two and four. An exception is perhaps the likelihood of holding a full-time position, which is lower for type seven than for all but the lowest type. They are also the most likely to have had health issues in the past, and second only to type one for having some other sort of break from work.

8 Conclusion

In this paper we develop and demonstrate a novel methodology for estimating treatment effects. In an application using novel French data on wages, we find results in line with previous work: formal training seems to have a small positive effect on wages. However, our main contribution is the method. We are able to estimate treatment effects conditional on latent types which control for selection into treatment on both observed and unobserved characteristics. Our approach is similar in spirit to differences-in-differences

Table 6: Analysis of individual and firm characteristics across types ($K = 7$)

	1	2	3	4	5	6	7
p_k	0.18	0.18	0.16	0.09	0.07	0.06	0.26
Individual characteristics							
Hourly wage (EUR, 2013)	9.04	11.45	14.50	19.16	24.40	32.78	14.60
Wage standard deviation	0.84	1.05	1.48	1.75	2.11	6.18	6.23
Full-time	0.79	0.93	0.95	0.95	0.96	0.95	0.86
Permanent contract (CDI)	0.92	0.97	0.98	0.99	0.99	0.98	0.91
PCS category							
PCS1	0.45	0.46	0.28	0.11	0.03	0.01	0.27
PCS2	0.42	0.21	0.12	0.05	0.03	0.02	0.19
PCS3	0.05	0.14	0.23	0.21	0.08	0.02	0.12
PCS4	0.07	0.15	0.20	0.15	0.08	0.05	0.14
PCS5	0.00	0.01	0.08	0.26	0.46	0.36	0.11
PCS6	0.01	0.02	0.06	0.20	0.29	0.46	0.15
Education							
Less than bac.	0.57	0.50	0.38	0.22	0.11	0.09	0.37
Bac. gen. or pro.	0.24	0.21	0.19	0.15	0.11	0.09	0.17
Bac+2 or more	0.19	0.28	0.43	0.62	0.78	0.82	0.45
Partner	0.63	0.74	0.77	0.81	0.83	0.87	0.73
Children	0.48	0.55	0.61	0.64	0.69	0.68	0.57
French	0.94	0.97	0.97	0.97	0.98	0.97	0.96
Female	0.44	0.29	0.25	0.22	0.17	0.14	0.35
Age							
Less than 30	0.29	0.17	0.11	0.06	0.03	0.01	0.21
30-40	0.23	0.27	0.28	0.28	0.27	0.17	0.27
40-50	0.28	0.33	0.37	0.37	0.40	0.39	0.28
Older than 50	0.20	0.23	0.24	0.29	0.30	0.44	0.25
Health issues (current)	0.12	0.11	0.08	0.05	0.03	0.02	0.16
LT contract (pre-'13)	0.61	0.62	0.63	0.70	0.70	0.72	0.61
ST contract (pre-'13)	0.50	0.52	0.45	0.39	0.33	0.26	0.43
Unemployed (pre-'13)	0.28	0.21	0.17	0.12	0.11	0.10	0.17
Health issues (pre-'13)	0.05	0.05	0.04	0.02	0.01	0.01	0.06
Other break (pre-'13)	0.14	0.08	0.05	0.05	0.05	0.03	0.09
Firm characteristics							
Firm size							
< 50	0.43	0.35	0.26	0.21	0.15	0.18	0.25
50-249	0.25	0.24	0.22	0.21	0.16	0.20	0.21
> 249	0.32	0.41	0.52	0.58	0.70	0.62	0.54
Sector							
Manufacturing	0.18	0.36	0.41	0.41	0.41	0.31	0.29
Services	0.79	0.60	0.56	0.57	0.56	0.66	0.68
CDD at firm*	12.64	9.35	7.48	5.31	5.75	8.36	8.50
Part-time at firm*	20.27	10.05	7.80	7.86	8.40	9.29	12.65
Individual incentives	0.52	0.63	0.72	0.76	0.83	0.79	0.70
Collective incentives	0.59	0.71	0.78	0.84	0.87	0.83	0.74
Outsource	0.26	0.34	0.41	0.46	0.50	0.43	0.40
HR department	0.78	0.85	0.89	0.91	0.94	0.93	0.89

* Not all individuals in our subsample gave a valid response to these questions. The values in these rows represent mean percentage of CDD or part-time employees at the firms of respondents assigned this type.

Notes: This table displays observable characteristics collected in the Defis survey by type, where type were assigned as previously using the maximal posterior probability. Unless stated otherwise, the values in the table are the probability of the given characteristic being true among respondents assigned to this group. Type seven is separated as rather than allocating individuals to this group by maximal posterior probability, we include only workers for whom $p_i(7) > 0.85$.

(which we compare it to here), but requires only an exclusion restriction (which are often motivated by economic theory, or even policy-relevant as in our application) rather than the usual common trends assumption. Our other important contribution is a non-parametric identification proof and a relatively parsimonious estimation strategy via the EM-algorithm.

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. van den 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," *The Quarterly Journal of Economics*, 113, 79–119.

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

Anonymous (2017): "Lifelong learning," *The Economist*, 9.

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

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

Autor, D. H., F. Levy, and R. J. Murnane (2003): "The skill content of recent technological change: An empirical exploration," *The Quarterly journal of economics*, 118, 1279–1333.

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.

Blundell, R., L. Dearden, and C. Meghir (1996): “The determinants and effects of work-related training in Britain,” Tech. rep.

Bonnal, L., D. Fougre, and A. Sérandon (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,” *The Review of Economics and Statistics*, 75, 164–170.

Brodaty, T., B. Crépon, and D. Fougre (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).

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.

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.

Crépon, 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.

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.

Ferracci, M. (2013): *Évaluer la formation professionnelle*, Sécuriser l’emploi, Presses de Sciences Po.

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

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

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

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

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

Heckman, J. J., R. J. Lalonde, and J. A. Smith (1999): "The economics and econometrics of active labor market programs," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 3 of *Handbook of Labor Economics*, chap. 31, 1865–2097.

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

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

Jamshidian, M. and R. I. Jennrich (2000): "Standard Errors for EM Estimation," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62, 257–270.

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," *The 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.

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.

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*, Elsevier, vol. 5 of *Handbook of the Economics of Education*, chap. 0, 479–652.

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.

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

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