Dynamic Treatment Effects of Job Training*

Jorge Rodríguez Fernando Saltiel Sergio Urzúa

December 15, 2017

Abstract

This paper investigates the dynamic returns to job training. We exploit unique administrative data combining job training records, matched employee-employer data on wages and employment, and pre-labor market ability measures from Chile. We estimate a dynamic-discrete model of training choices and earnings where workers self-select into training based on observed and unobserved characteristics. We test and reject the conditions for fixed-effect estimators to recover average effects of training. We document static and dynamic effects, finding substantial heterogeneity in the estimated returns. Low-ability individuals are less likely to participate in a training course, but have higher labor market gains from training.

Keywords: Dynamic Treatment Effects, Training, Unobserved Heterogeneity.

* *Corresponding author*: Sergio Urzúa, Department of Economics, University of Maryland and NBER; email: urzua@econ.umd.edu. Jorge Rodriguez, Department of Economics, University of Chicago; email: jorger@uchicago.edu. Fernando Saltiel, Department of Economics, University of Maryland; email: saltiel@econ.umd.edu. . We are indebted to seminar participants at LACEA-LAMES, RIDGE's Impact Evaluation of Labor Market Policies Conference, *Banco de Desarrollo de America Latina*, PUC-Rio, and PUC-Chile.

1 Introduction

In a context of rapid technological change, occupations constantly change their skill requirements.[1] However, workers may be endowed with a different set of competencies vis-á -vis those required in the workplace.[2] To mitigate such gaps, firms and workers invest in on- or off-the-job training. In fact, as occupations today require a multitude of skills, workers could find it necessary to participate in training courses on multiple occasions.

The aim of this paper is to estimate the returns to training in the context of dynamic training choices. The existence of dynamic program participation presents an inherent challenge for estimating these returns, as part of the return to training in the present period may reflect an increase in the returns to future training events. Furthermore, forward-looking workers may consider the option value of training when making their participation decisions. In this context, we first examine whether commonly-used reduced-form estimators are able to recover relevant parameters, such as the average treatment effect. Similarly, we explore whether such strategies can capture the returns to program participation across multiple years. To deal with these questions, we extend a static choice framework to consider a model where training decisions take place across various periods and outcomes can vary freely across training choices. Using our model, we show that reduced-form estimators are generally not informative about the short-and long-run effects of job training. More generally, our findings illustrate the importance of accounting for dynamic program participation and returns when evaluating the gains from human capital investments.

We present a tractable dynamic-discrete choice model that captures a variety of static and dynamic treatment effects.[3] In the model, a worker must decide whether or not to take part in a training course across multiple periods. For a given training history, and conditional on firm characteristics, the agent decides to be trained if the associated net benefits are positive.[4] Individual choices and outcomes depend on observed characteristics as well as on unobserved heterogeneity,

[1]See Autor et al. (2003), SpitzOener (2006), and Ingram and Neumann (2006), Acemoglu and Autor (2011), and Sanders and Taber (2012) for a literature on the changing returns to specific skills.

[2]Guvenen et al. (2015), Buhrmann (2017), Saltiel et al. (2018) discuss the importance of matching in the labor market.

[3]Our model extends Heckman and Navarro (2007) and Heckman et al. (2016) framework to the particular characteristics of our data.

[4]Besides the index threshold-crossing property that determines the decision process, we do not model preferences and budget sets, thereby abstracting from strong assumptions about behavior and uncertainty—ubiquitous elements in the structural dynamic literature (Keane et al., 2011). Instead, we approximate choices and outcomes processes within a joint framework, which accounts for self-selection into program participation at each decision node.

which we interpret as the initial stock of skills. Using a measurement system of test scores, we are able to nonparametrically identify the distribution of unobserved latent skills and use it to identify the joint distribution of counterfactual earnings across potential training choices.

Our discrete choice framework can be applied in any setting in which agents can self-select into program participation in multiple occasions. In this paper, we focus on a large scale training program in Chile, "*Franquicia Tributaria*" (FT), which fully subsidizes training courses at off-site providers for workers who are employed in a formal sector firm. Hundreds of different courses can be subsidized under FT. In the program, a worker can participate in training choices on multiple occasions, and a significant number of workers do so. We take advantage of administrative data on job training records for the population of workers and combine it with matched employee-employer data on wage and employment outcomes, allowing us to construct panel data including workers' participation histories and their respective wages. We further combine this data with information on measures of workers' pre-labor market abilities coming from college admission test scores. We focus on the returns of training for first-time labor market entrants.[5]

We document two main results. First, we evaluate the approach followed by the traditional literature, which estimates average treatment effects by implementing some variant of a fixed-effects estimator. Within a context of a general model of counterfactual outcomes across all possible training histories, we initially document the assumptions required for fixed effects estimators to recover average treatment effects. Using our estimated discrete choice model, we test and reject these assumptions, thereby concluding that fixed-effect estimators are not able to recover average treatment effects in this setting.

In a second set of results, we exploit our model to document static and dynamic treatment effects of training. We decompose the effects of training in one period into direct (short-term) effects and continuation values. While direct effects are positive on average, continuation values are negative, though small in magnitude. Thus, we find evidence of negative dynamic complementarities: training in one period reduces the returns to training in a second period. Both direct and continuation values vary by unobserved ability. In particular, direct effects and continuation values are positive

[5] As we observe the entire training and employment history of these workers in their first two years in the labor force, we do not confound the returns to this program with existing stock of previous training programs. Furthermore, as argued by Mincer (1974) firms may find it especially valuable to train younger workers as there is a longer time period over which they can capture the returns to training.

for low-ability workers, although they participate less often in training courses than high-ability individuals. This result may be explained by the fact that FT courses often are often focused on basic skill development, and thus provide positive returns for low-ability workers.

Our paper contributes to a long literature on the returns to training courses. The inherent challenge in estimating the returns to job training stems from potential self-selection into training (Heckman et al., 1998). As a result, various papers in the literature have relied on individual fixed-effect estimators to account for unobserved individual heterogeneity. In the meta-analysis of Card et al. (2010), 51% of the studies exploit the longitudinal structure of the data.[6] Moreover, some papers have found that versions of the standard fixed-effect estimator can effectively remove selection bias (Heckman et al., 1999, 1998). However, to our knowledge, there are no papers evaluating what fixed-effect estimators identify in a context where the returns to training depend on the history of training choices.

An important portion of the training literature examines the effect of training programs incorporating an experimental component (Attanasio et al., 2011; Card et al., 2011; Alzúa et al., 2014; Attanasio et al., 2015; Ibarrarán et al., 2015). While experimental estimators may reliably identify causal effects, they generally suffer from at least three limitations. First, sample sizes tend to be small, thereby limiting the external validity of empirical estimates. Second, these papers have often analyzed short-term outcomes. Third, experimental studies do not directly identify the mechanisms behind the effects of training. Our framework allows us to overcome these limitations. We examine the impact of training on medium-term outcomes and estimate the dynamic returns to training. We additionally identify the mechanisms driving the dynamic effects by decomposing the returns into direct effects and continuation values.

Finally, a number of papers have previously estimated dynamic returns to training for unemployed workers (Abbring and van den Berg (2003); Fredriksson and Johansson (2008); Fitzenberger et al. (2016)). Nevertheless, little is known about the dynamic treatment effects of job training for employed workers, making our paper a first in this context. This question is of particular interest to labor economists, as existing theoretical models show that both firms and workers should benefit from programs which provide employees with a broad set of skills required to perform the job (Acemoglu and Pischke, 1999). Our discrete choice model, presented below, overcomes potential

[6] See also Ashenfelter (1978), Lynch (1992), Booth (1993), Veum (1997), Lengermann (1999).

concerns about the no-anticipation assumption and allows us to make a significant contribution to this literature by analyzing the returns to multiple training periods for employed workers using administrative data.

The paper is organized as follows. Section 2 presents our model of sequential training participation with unobserved heterogeneity and shows the limitation behind reduced form approaches. Section 3 describes the institutional setup, data and sample of our program of interest. Section 4 presents evidence on the model's fit and on selection into training on ability. In section 5, we define the static and dynamic treatment parameters and present our estimated training effects. We conclude in Section 6.

2 Conceptual Framework

2.1 A dynamic model of training decisions and earnings

In this section, we introduce a dynamic Roy model to describe the dynamics of training decisions and labor market earnings. Our model follows Heckman and Navarro (2007) and Heckman et al. (2016) (albeit with a few accommodations). In the model, an agent, after entering the labor market for the first time and taking into account the characteristics of the firm, must decide on whether to be trained in each period. There are T periods, where $t = 1$ denotes the period at which she enters the labor market and T the period in which she exits. The agent is allowed to be trained as many times as she desires, and we focus on her extensive margin decision.[7] In any period t, potential earnings depend on her current training decision as well as on the entire history of past training choices. The agent chooses training at period t taking into account past training decisions and, possibly, current and future potential outcomes. As will be seen, we write "possibly" since we do not need to specify preferences and future expectations formation.

We model the dynamic training decision as a tree of sequential binary decisions, where the individual chooses training in each stage $t \in \mathcal{T} \equiv \{1, ..., T\}$. Figure 1 depicts this decision tree for the special case where $T = 2$. An element in the set of possible training histories encompassing all

[7] At this point, we depart from Heckman and Navarro (2007) and Heckman et al. (2016). In their model, one particular education path restrains the choice set (for example, getting a GED implies that the individual cannot attain the high school graduate node). In contrast, our model allows for the option of training independently of previous decisions.

periods up through period $t-1$ is given by $h_t \in \mathcal{H}_t$. For instance, in period $t = 2$, $\mathcal{H}_2 = \{0, 1\}$, where $h_2 = 1$ if the agent was in $t = 1$ trained and 0 otherwise. Likewise, $\mathcal{H}_3 = \{(0,0), (1,0), (1,0), (1,1)\}$, where each element $h_t \equiv (i, j)$ denotes the training decision in period $t = 1$ (first component) and $t = 2$ (second component).

At each stage, agent i solves a simple benefit-cost analysis to decide whether she will be trained given her employer's characteristics. Let $D_{it}(h_t)$ denote the training decision variable at period t, with history $h_t \in \mathcal{H}_t$. Her optimal choice is given by:

$$D_{it}(h_t) = \begin{cases} 1 & \text{if } I_{it}(h_t) \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad h_t \in \mathcal{H}_t, t \in \mathcal{T} \tag{1}$$

where $I_{it}(h_t)$ denotes the value of training at t for a given history $h_t \in \mathcal{H}_t$. We specify $I_{it}(h)$ in subsection 2.3.

The agent progresses through each node after making training choices, and for each possible choice and training history, there is an associated labor market outcome. Let $Y_{it}(j, h_t)$ be an outcome (such as earnings or employment) for a training decision $j \in \{0, 1\}$ made by worker i at time t with history h_t. The counterfactual outcome follows:

$$Y_{it}(j, h_t) = \mu^Y(j, h_t) + \nu^Y_{it}(j, h_t), \tag{2}$$

where $\mu^Y(j, h_t)$ is a parameter that varies by the current training choice and training history, and generally includes workers' observable characteristics.[8] As seen in equations (5) and (6), we allow for workers' prior training decisions to affect observed outcomes at time t. $\nu^Y_{it}(j, h_t)$ is latent productivity, which is unobserved by the econometrician. This unobserved component is separable in a productivity component—known by the agent—and an idiosyncratic shock as follows:

$$\nu^Y_{it}(j, h_t) = \lambda^Y(j, h_t)\theta_i + \epsilon^Y_{it}(j, h_t), \tag{3}$$

[8] In fact, we could more generally assume that this parameter varies by i, $\mu^Y(j, h_t, \boldsymbol{X}_i)$, where $\boldsymbol{X}_i$ is a vector of observables. However, our results would hold by adding these control variables in the regressions while making notation more cumbersome for the purposes of this Section. In Section 3.4, we assume a linear specification for $\mu^Y(j, h_t, \boldsymbol{X}_i)$ to take the dynamic model of training to a tractable, econometric framework.

where $\epsilon_{it}^{Y}(j,h_t)$ is unobserved by both the econometrician and the agent and where θ_i is a fixed, latent skills endowment observed by the agent who can thus act upon it. We assume $E[\theta_i]=0$.

The model of choices and counterfactual outcomes consists of equations (2) and (3). The model is general enough to result in rich dynamics in the returns to training; in both equations, the returns to the observed and unobserved characteristics are allowed to vary by past training histories and current choices. We next explore the usefulness of reduced-form estimators to capture relevant treatment effects within this model.

2.2 Reduced-Form Estimation

We begin this section by introducing one parameter of interest in the training literature, the average treatment effect. Following the logic of the previous section, the impact of training for an individual i at period t for a given history h equals $Y_{it}(1,h_t)-Y_{it}(0,h_t)$. Let $H_{it}(h_t)$ be an indicator variable that equals 1 if individual i in period t followed training history h and 0 otherwise. The overall average of these individual treatment effects is defined as:

$$\begin{aligned}ATE &\equiv E\left[\sum_{h\in\mathcal{H}_t} H_{it}(h_t)\left(Y_{it}(1,h_t)-Y_{it}(0,h_t)\right)\right]\\ &= E\left[\sum_{h\in\mathcal{H}_t} H_{it}(h_t)\left(\mu^Y(1,h_t)-\mu^Y(0,h_t)+(\lambda^Y(1,h_t)-\lambda^Y(0,h_t))\theta_i\right)\right], \end{aligned} \tag{4}$$

where the expected value operator integrates with respect to i and t. Hence, the overall average treatment effect is a weighted average of individual treatment effects across periods, at different potential training histories.

In a longitudinal data set-up, the analyst's goal is to identify (4) using observed data (Y_{it},D_{it}), where D_{it} and Y_{it} represent the observed training indicator and outcome variable. Following the standard switching regression model, we can express them as functions of underlying potential outcomes and choices. Observed variables are given by:

$$D_{it}\equiv\sum_{h_t\in\mathcal{H}_t} H_{it}(h_t)D_{it}(h_t), \tag{5}$$

$$Y_{it}\equiv\sum_{h_t\in\mathcal{H}_t} H_{it}\left[D_{it}(h_t)Y_{it}(1,h_t)+(1-D_{it}(h_t))Y_{it}(0,h_t)\right] \tag{6}$$

The empirical literature examining the impact of job training first proposes linear regression models of the form:

$$Y_{it} = \pi_0 + \pi_1 D_{it} + \xi_{it} \qquad \text{for } i = 1, ..., N \text{ and } t = 1, ..., T \tag{7}$$

where ξ_{it} is an error term. Ordinary least squares in this context identifies:

$$\delta^{OLS} \equiv \frac{Cov(Y_{it}, D_{it})}{Var(D_{it})} = E[Y_{it}|D_{it} = 1] - E[Y_{it}|D_{it} = 0]$$

If the data generating process follows our dynamic model, then potential self-selection into training results in a correlation between ξ_{it} and D_{it} (Ashenfelter and Card (1985)). To see how self-selection affects the reduced-form estimate, let:

$$\mu^Y(j) \equiv \sum_{h \in \mathcal{H}_t} H_{it}(h_t)\mu^Y(j, h_t) \quad \lambda^Y(j) \equiv \sum_{h_t \in \mathcal{H}_t} H_{it}(h_t)\lambda^Y(j, h_t) \quad \epsilon^Y_{it}(j) \equiv \sum_{h_t \in \mathcal{H}_t} H_{it}(h_t)\epsilon^Y_{it}(j, h_t)$$

for $j \in \{0, 1\}$. Using the definition of D_{it} and Y_{it} (equations 5 and 6), and summing observed and unobserved parameters across training histories, we have the following equation:

$$Y_{it} = \mu^Y(0) + D_{it}(\mu^Y(1) - \mu^Y(0)) + \xi_{it}$$

where the unobserved part of the equation is:

$$\xi_{it} \equiv (\lambda^Y(0)\theta_i + \epsilon^Y_{it}(0)) + D_{it}(\epsilon^Y_{it}(1) - \epsilon^Y_{it}(0)) + (\lambda^Y(1) - \lambda^Y(0))\theta_i$$

Thus, if net benefits of training ($I_{it}(h_t)$ in equation 1) depend on the unobserved latent ability endowment, then the OLS estimator of π_1 (equation 7) is inconsistent. Since the inconsistency is originated because the analyst does not observe θ_i—and, thus, it cannot control for it—, one commonly-used approach is to take advantage of the longitudinal nature of the model and add an individual fixed-effect parameter:

$$Y_{it} = \pi_0 + \pi_1 D_{it} + u_i + v_{it}. \tag{8}$$

One way of estimating (8) is by taking First Differences (FD). Since we observe (Y_{it}, D_{it}) for various periods, we could estimate by OLS on:

$$\Delta Y_{it} = \pi_1 \Delta D_{it} + \Delta v_{it},$$

where the fixed effect has been eliminated.[9] In Appendix A, we show the FD estimator yields:

$$\begin{aligned}
\delta^{FD} &\equiv \frac{Cov(\Delta Y_{it}, \Delta D_{it})}{Var(\Delta D_{it})} \\
&= 1/2 \times E[\Delta Y_{it} | \Delta D_{it} = 1] - 1/2 \times E[\Delta Y_{it} | \Delta D_{it} = -1] \qquad (9)
\end{aligned}$$

Note that if training decisions follow our proposed data generating process, from the definition of Y_{it}, we can get an equivalent expression using the underlying structural elements, but this time bringing the elements associated with θ_i in the "observables" part:

$$\begin{aligned}
Y_{it} &= \overbrace{\mu^Y(0)}^{\pi_0} + D_{it} \overbrace{(\mu^Y(1) - \mu^Y(0))}^{\pi_1} + \overbrace{[(\lambda^Y(1) - \lambda^Y(0))\theta_i + \lambda^Y(0)\theta_i]}^{u_i} + v_{it}, \\
v_{it} &\equiv \epsilon^Y_{it}(0) + D_{it}(\epsilon^Y_{it}(1) - \epsilon^Y_{it}(0)).
\end{aligned}$$

By controlling for the unobserved coefficient u_i, we can therefore recover consistent estimates of π_1. However, which treatment parameter is the FD estimator recovering? We next show that this approach can only identify the average treatment effect specified in equation (4) under assumptions that imply ignoring the basic dynamics underlying training choices.

Consider the following assumptions:

Assumption 1. *In equation (2), $\mu^Y(1, h_t) - \mu^Y(0, h_t) = \pi_1$ for $h_t \in \mathcal{H}_t$ and $t \in \mathcal{T}$.*

Assumption 2. *In equation (2), $\lambda^Y(1, h_t) - \lambda^Y(0, h_t) = 0$ for $h_t \in \mathcal{H}_t$ and $t \in \mathcal{T}$.*

Assumption 1 defines the return to on-the-job training. It restricts this return to be constant for all training histories, as well as in across all time periods $t \in \mathcal{T}$. Assumption 2 rules out any potential gains to treatment for individuals with different levels of unobserved heterogeneity,

[9] As is shown in several papers in the literature, when $T = 2$, the fixed-effect estimator is equivalent to the first-differences estimator. In this paper, we focus on the first-differences estimator, but the results are equivalent in the fixed effect context.

thereby ruling out the possibility that individuals with higher levels of unobserved ability may enjoy larger returns to training. Under these assumptions, we can show the following.

Proposition 1. *Suppose outcomes are determined by equation (2) and Assumptions 1 and 2. Then the FD estimator from equation (9) follows:*

$$\delta^{FD} = \pi_1 = \mu^Y(1, h_t) - \mu^Y(0, h_t) \qquad \textit{for } h_t \in \mathcal{H}_t, t \in \mathcal{T}$$

Proof. See Appendix B. □

Under assumptions 1 and 2, Proposition 1 shows that the FD estimator recovers an average treatment effect which is constant in time and across histories. As a result, these assumptions not only impose strong restrictions within periods, but also across labor market and training histories; Assumptions 1 and 2 imply that the returns to training are equivalent for workers trained at time t with training histories $h \in \mathcal{H}_t$ and $h'' \in \mathcal{H}_t$ as well as for workers trained at time $t-1$ with histories $h' \in \mathcal{H}_{t-1}$. Furthermore, these assumptions imply absence of complementarities in the human capital accumulation process, which is a particularly strong assumption in the context of skill development in the labor market (Mincer 1974).

Following our dynamic model of program participation, we can define dynamic treatment effects. In the context of our model of training choices, dynamic treatment effects can be of interest as they allow capturing potential complementarities in the returns to training. For instance, we may be interested in estimating the effect of training for a worker who has received training at time t and $t-1$ relative to a counter-factual history with no training in either period. In this context, we can define the following dynamic treatment effect parameter:

$$E[Y_{it}(1, (1, h')) - Y_{it}(0, (0, h'))], \qquad h' \in \mathcal{H}_{t-1} \tag{10}$$

Are reduced form strategies able to recover dynamic treatment effects? One option is to use OLS to compute $E[Y_{it} \mid D_{it} = 1, D_{it-1} = 1] - E[Y_{it} \mid D_{it} = 0, D_{it-1} = 0]$. But this estimation procedure does not account for self-selection into training (in t and $t-1$), yielding an inconsistent estimator in the context of our dynamic model. Furthermore, we cannot use the FD estimator since it requires using the sample of workers who have changed their participation decision in periods t and $t-1$.

By not incorporating information from workers who have kept their training decisions constant in time, the FD estimator is unable to recover the dynamic effect of equation (10). Therefore, a fixed effects approach cannot identify dynamic complementarities. We next present an econometric framework that allows us to recover relevant static and dynamic treatment effect parameters as well as heterogeneous impacts of job training across training histories and time.

2.3 Dynamic Setting

In this sub-section, we present our dynamic econometric framework of training and earnings. We extend the decision model of Section 2.1 to transform it into a econometric, tractable model. Potential earnings at each node depend on a set of observed and unobserved characteristics. Instead of fully specifying individual's preferences and economic constraints, we approximate the individual value of training at each node using a linear specification that depends on observed and unobserved characteristics as well. In this way, the econometric model refrains from relying on functional form and distributional assumptions. Our model can be nonparametrically identified, similarly to the LATE framework.[10] On the other hand, unlike the LATE approach, we are able to identify different margins of treatment.

As mentioned above, we approximate the decision process by defining worker i's value of training, $I_{it}(h_t)$, using the following linear-in-the-parameters equation:

$$I_{it}(h_t) = X_i^I \beta^I(h_t) + \theta_i \lambda^I(h_t) + \varepsilon_{it}^I(h_t) \qquad h_t \in \mathcal{H},\ t \in \mathcal{T}, \tag{11}$$

where X_i^I is a vector of individual characteristics (observed by the econometrician and the agent), θ_i represents a latent skill endowment known by the agent but not the econometrician, and $\varepsilon_{it}^I(h_t)$ is an unobserved measurement error term. This specification can accommodate any type of preferences structure (Heckman et al., 2016). In particular, it can be considered as an approximation of the value function of a forward-looking agent looking to maximize the present value of earnings' streams.

As discussed above, each training choice j and history h_t have associated a potential labor

[10] Rust (1994) and Magnac and Thesmar (2002) show that a class of dynamic-discrete choice models cannot be nonparametrically identified

market outcome, captured as follows:

$$Y_{it}(j,h_t) = X_i^Y \beta^Y(j,h_t) + \theta_i \lambda^Y(j,h_t) + \varepsilon_{it}^Y(j,h_t) \qquad h_t \in \mathcal{H}_t,\ t \in \mathcal{T},\ j \in \{0,1\}, \tag{12}$$

where X_i are individual characteristics, θ_i represents the latent ability endowment, and $\varepsilon_{it}^Y(j,h_t)$ is an idiosyncratic shock to earnings.

As equations (11) and (12) imply, a common (and constant) factor, θ_i, drives the endogeneity of choices and outcomes. Following a large body of literature on static and dynamic treatment effects, we extract this factor using a measurement system.[11] Using such a system allows us to interpret this common factor as an unobserved ability component. The rest of unobserved components of the model are independent across choice and outcomes equations, and across time and training histories.[12] Thus, conditional on X_i^Y and X_i^I, θ generates all cross-correlations of outcomes and choices. θ generates the so-called "ability bias" (Card, 1999): high-ability individuals can have higher returns from training than low-ability workers. As discussed above, since ability is unobserved by the econometrician, OLS estimates of equation (12) are inconsistent.

We do not observe θ, but approximate it by a set of variables measuring an initial ability endowment (prior to the training decision). Let T_k represent a measure of ability, for $k \in \mathcal{K} \equiv \{1,2,3\}$. We use three measures of pre-labor market ability, including high school GPA, a worker's score on a high-school exit exam in mathematics and language, along with a measure of her labor market sklils, represented by his/her first monthly salary upon labor market entry. We state the following linear factor model:

$$T_{ik} = X_{ik}^T \beta_k^T + \theta_i \lambda_k^T + \xi_{ik} \qquad k \in \mathcal{K}, \tag{13}$$

where X_{ik} is a vector of exogenous control variables and $\varepsilon_{it}^I(h) \perp \xi_{ik}$ for all $(h,k) \in \mathcal{H} \times \mathcal{K}$, $\varepsilon_{it}^Y(j,h) \perp \varepsilon_{ik}$ for all $(t,j,k) \in \mathcal{T} \times \{0,1\} \times \mathcal{K}$, and $\varepsilon_k \perp \varepsilon_v$ for all $k, v \in \mathcal{K}$. Finally, we assume that unobserved ability is orthogonal to $\mathbf{X}_i \equiv (X_i^Y, X_i^I, \{X_{ik}^T\}_{k \in \mathcal{K}})$ (the vector that contains all

[11] Carneiro et al. (2003), Hansen et al. (2004), Heckman et al. (2006), Heckman and Navarro (2007), Heckman et al. (2013), Attanasio et al. (2015), Agostinelli and Wiswall (2016a), Agostinelli and Wiswall (2016b), and Heckman et al. (2016) are just a few examples.

[12] Formally, we assume that $\varepsilon_{it}^I(h) \perp \varepsilon_{it}^I(h')$ for all $(t,h) \in \mathcal{T} \times \mathcal{H}$, $\varepsilon_{it}^I(h) \perp \varepsilon_{it}^Y(j,h)$ for all $(t,j,h) \in \mathcal{T} \times \{0,1\} \times \mathcal{H}$, and $\varepsilon_{it}^Y(j,h) \perp \varepsilon_{it'}^Y(j',h')$ for all $(t,j,h) \in \mathcal{T} \times \{0,1\} \times \mathcal{H}$

observable variables).

Heckman and Navarro (2007) show identification of a similar dynamic discrete choice model.[13] In our setting, the goal is to identify a joint distribution of counterfactual earning outcomes for all potential training decisions and labor market histories. There are three key steps for identification of a joint distribution of counterfactuals. First, we need independent variation in the choice and outcome equations to apply an identification-at-infinity argument. In the linear case, this is achieved by using instruments with enough support. In our context, we use the average per worker number of program training hours used by the worker's employer in the previous year as an instrumental variable. Second, we invoke that, conditional on θ and $\mathbf{X}$, all choices and outcomes are independent—a "matching" assumption. Third, we identify the distribution of θ using the measurement system. In this step, one can show that the distribution of the factor is nonparametrically identified up to a normalization (which usually takes the form of setting one factor loading $\alpha_k = 1$). By identifying the joint distribution of counterfactual outcomes, the distribution of all kinds of treatment effects (ATE, TT, TUT, and so on) are also identified. In this paper, we exploit this identification result when exploring returns to training for different decisions margins.

Estimation

We now present the likelihood function and some details about our estimation procedure. We adopt a flexible approach to estimate the unobserved ability's distribution function $F(\theta)$. Specifically, we assume the factor is drawn from a mixture of normal distributions:

$$\theta \sim \rho_1 N(\tau_1, \sigma_1^2) + \rho_2 N(\tau_2, \sigma_2^2) + \rho_3 N(\tau_3, \sigma_3^2)$$

To define individual the sample likelihood, collect all observable in the vector $\mathbf{X}_i \equiv (X_i^Y, X_i^I, \{X_{ik}^T\}_{k \in \mathcal{K}})$ and define the set of structural parameters as Ψ. Given our independence assumptions, the likelihood for a set of $\mathcal{I}$ individuals is given by:

$$\mathcal{L}(\Psi \mid \mathbf{X}, Y) = \prod_{i \in \mathcal{I}} \left[\int_\theta \prod_{k \in \mathcal{K}} f(T_{ik} \mid X_{ik}^T, \theta) \prod_{t \in \mathcal{T}} \prod_{h_t \in \mathcal{H}_t} \left\{ f(Y_{it} \mid X_i^Y, \theta) Pr(H_t = 1 \mid X_i^I, \theta) \right\}^{H_t} dF(\theta) \right],$$

[13] Their result follows applying Theorems 1 and 2 to any particular node. See also Heckman et al. (2016).

where we assume that idiosyncratic shocks in the choice process (equation 11), earnings regression (equation 12), and measurement system (equation 13) follow independent normal distributions. Given the numeric complexity in estimating the likelihood, we estimate the model by Markov Chain Monte Carlo (MCMC). Using the estimated model, we simulate 200,000 observations. Table 1 shows the variables used in the implementation of the model, for the wage equation, the participation probits and the measurement system. We next describe the institutional setting in which we carry out our empirical strategy.

3 Training Program, Data and Sample

In this paper, we take advantage of a nationwide funding scheme for job training programs called *Franquicia Tributaria* (FT) in Chile. The program covers a significant portion of the formal sector labor force: on average, twelve percent of formal sector workers are trained under this program each year. FT fully subsidizes firms for the cost of training its employees in courses offered by private providers, called Technical Training Institutes (TTI). There are over 15,000 providers in the market and the market share of each provider remains low, yielding a highly competitive training system. Moreover, providers offer a variety of courses, including industry-specific courses and courses providing general training. While all formal sector firms are theoretically eligible for the program, in practice, only firms with at least ten employees are able to take advantage of the full subsidy. Moreover, as the government provides a tax credit of up to 1 percent of the annual payroll, the majority of employees at eligible firms are able to become trained each year if they so desire. Finally, workers and firms are able to choose the courses they deem relevant for the worker's training, yielding a system that could theoretically allow workers to develop the necessary bundle of skills required to adequately perform their jobs. Courses are largely offered after-work hours and on weekend such that they do not interfere with regular work schedules.

As discussed above, we rely on administrative data for workers' training histories as well as matched employee-employer data. We construct workers' training histories by observing their participation in FT-subsidized courses from 1998 through 2010. For each training stint, we identify the length of the program, total hours per course, and certain characteristics of the associated providers. We additionally use data from the Chilean Unemployment Insurance system, which registers the firm of employment and the monthly record of earnings for all workers with formal labor market contracts. The database also includes information on workers' observable characteristics, such as gender, age, geographic location, education, parental education, family size and family income. For a number high school graduates between 2000 and 2007, we also observe relevant measures of pre-labor market ability, including their scores on university entrance exams (PSU) in math and language skills. As a result, we are able to estimate the impact of training programs on a large number of trainees.

In this paper, we focus our attention on the returns to multiple job training courses for young workers who are first-time labor entrants. We observe the first year of employment for this group of workers and track their employment and training histories over time.[14] For model tractability, we restrict our analysis to training stints during their first two years in the labor force and examine extensive margin training decisions, such that each year, the worker can either be trained or not. As a result, workers are trained at most two times during our period of interest. While the training decision is made jointly by the firm and the worker, we restrict our sample to a group of workers who are able to take at least one course in each year, as our analysis of worker self-selection into training requires workers to be able to partake in the courses each year if they so desire.[15] In our empirical analysis, we select a random sample of 5,000 workers who meet the criteria stated above. These workers are on average 21 years old when they first enter the formal labor force.

Table 2 presents summary statistics for the group of eligible program participants. The sample is leans mostly female and we report the age at the time of test-taking, which is in line with the average for the country. Furthermore, the average PSU score in math and language is above the average in Chile, reflecting the fact that higher ability workers participate in the formal sector.

[14] Arellano and Honore (2001) have outlined the assumptions required to achieve identification in a context with state dependence and unobserved heterogeneity where participants are observed after treatment has started. This is not a direct concern in our set-up as we observe workers upon labor market entry, who have not yet participated in training. Ba et al. (2017) similarly highlight potential problems which may arise in a setting in which participants do not immediately undertake training participation, but instead wait to do so in future periods. To address this concern, we restrict our sample to a group of workers who are eligible to be trained in every period.

[15] The first restriction is to limit the sample to workers employed in firms which use the *Franquicia Tributaria* program. We additionally require workers to be employed in the formal sector for at least nine months in each of their first two years in the labor force such that they are able to self-select into training courses. In this paper, we do not explore the effects of program participation on the employment margin, but instead focus on the effect on wages by focusing on workers with strong labor market attachments. While restricting our attention overlooks the effect of training on workers' formal sector employment, we note that our focus on wages makes our paper comparable in scope to the existing literature on on-the-job training programs. In future work, we plan on extending our framework to account for employment effects.

Finally, the average monthly salary in the first quarter of the third year in the labor force for this group equals 749 dollars. Table 3 shows summary statistics by training choice group during their first two years in the labor force. The group is denoted by $\{h, h\}$, where $h \in \{0, 1\}$, such that $\{1, 1\}$ represents the group of always-trained workers and $\{0, 0\}$ the never trained group. The never trained group is by far the largest in our sample, representing 61 percent of all individuals. On average, this group leans slightly female and has a lower PSU and GPA average than workers in other groups. Most importantly, this group has the lowest average monthly salary in their third year in the labor market. For the intermediate groups, we examine whether there are significant differences among those groups receiving the same amount of total training. An interesting pattern seems to emerge both for the groups with one total training stint ($\{0, 1\}$ and $\{1, 0\}$): those who receive the training at a later point in their careers earn higher salaries in their third year, despite similar ability levels. These inter-temporal differences across the groups with similar overall training histories highlight the importance of the dynamic nature of our research question. In fact, early-trainees may earn lower wages vis-á -vis later trainees due to skill depreciation. Finally, as expected, the always trained group are the highest ability workers and subsequently earn the highest wages in the third year. The ability differences across groups as well as the average group differences given training timing highlight the importance of our structural model for correctly answering this research question.

4 Results

Model Estimates

Table 4 shows the factor loadings on the four measurement variables. Males are more likely to score higher on the math PSU test and earn higher initial wages, but score lower on the language one and have lower high school GPAs. Younger students at the time of test-taking are less likely to perform well in the high school graduation exam. We similarly find that older workers upon labor market entry are more likely to earn higher wages. Lastly, we find that students coming from more educated families and those whose parents are more strongly attached to the labor force perform better across the three test measures.

In Figure 2, we first present the unconditional distribution of the unobserved ability factor and the estimated parameters from the three mixture of normals which generate this distribution. The factor does not appear to be distributed normally, highlighting the importance of allowing for non-parametric estimation of this unobserved characteristic.

Model Validity

To examine the validity of the structural model, we conduct various goodness of fit tests. We first contrast workers' actual training decisions against those simulated in this model. We present the results in Figure 3, which shows that the model accurately predicts the share of workers who choose each of the four training sequences. We also contrast actual and simulated average wages in each of the six nodes of the decision tree. Table 5 presents the means and standard deviations of average wages along with the p-values of a χ^2 test of the equality of means, which shows that we are unable to reject the null hypothesis of the equality of actual and simulated wages in any of the six nodes.

To analyze the relative importance of cognitive ability in explaining the three test score measures used in this paper, we present a variance decomposition of the measurement system in Figure 4. The results show that observed variables are not major contributors, which is not surprising as we only include family observables, age, gender and family employment and composition variables in the measurement system. On the other hand, we find that the ability factor that we measure loads largely on the initial salary, rather than on the three test-based ability measures. As a result, the unobserved ability measure explains 63% of the variance in the initial salary, but only 13%, 8% and 8% of the math PSU and verbal PSU scores and high school GPA, respectively. We note that including a worker's initial salary as part of the measurement system involves an inherent trade-off: while it reduces the share of the variance in test scores explained by the unobserved factor, it allows us to better capture a worker's unobserved ability at the time of under-taking training decisions by incorporating a measure of his/her initial productivity in the labor market, which may reflect different abilities than high school examination exams.[16]

Reduced-form estimates

In Table 6, we present the estimated impact of job training in our sample following both a

[16]We note that our results are robust to the inclusion of different ability measures in our measurement system.

least squares estimation and a first differences approach.[17] Confirming the results presented by Frazis and Loewenstein (2007), we find that an ordinary least squares approach leads to large estimated impacts of job training, but once we control for unobserved heterogeneity through an individual fixed-effect, the estimated impact of job training is significantly attenuated and falls to 1.6 percent. In Column (4), we include interactions with the observed test score measures and find that the estimated return to training does not change. We note that despite the inclusion of these interaction terms, reduced form strategies cannot cannot capture self-selection into training in each of the decision nodes.

To examine the importance of self-selection into training, we focus on the role of ability in determining individual training participation at each decision node, for which we rely on the simulated distribution of the factor. In Figure 5, we compare the ability distribution for workers who choose to get trained both years against the distribution for the never-trained group, which shows that the ability distribution first order stochastically dominates that of the latter group. This result highlights the importance of controlling for worker self-selection into training to correctly estimate the relevant treatment parameters. Given the role of selection into training and considering the possibility that returns may vary across nodes and training histories, we next examine whether the first difference estimator can in fact recover average treatment effects of job training.

5 Treatment Effects

5.1 The first-difference and the Average Treatment Effect

We begin the presentation of our structural model results by examining whether assumptions (1) and (2) are met in our context. Since our model allows us to recover all relevant parameters in the switching regression framework presented in Section 4, we directly test the validity of these assumptions. As noted above, if these assumptions are not met, a first differences estimator will not be able to recover any treatment effect parameters.

Table 7 (Panel A) presents the parameters associated with assumption (1). In our setting, assumption (1) requires that the coefficients associated to X_i^Y (observable in the earnings equations)

[17] Y_{it} represents average monthly wages in the first quarter of the year after the training event. For notational simplicity, we refer to the outcome variable as contemporaneous with the training event at time t.

are such that $\mu^Y(1, h_0) - \mu^Y(0, h_0) = \mu^Y(1, 1) - \mu^Y(0, 1) = \mu^Y(1, 0) - \mu^Y(0, 0) = \pi_1$ for $h \in \mathcal{H}_t$, where μ^Y are coefficients associated with each training choice. We test the equality of these three variables using a χ^2 test where the null hypothesis is the three parameters equal each other. Since the three parameters are statistically different from each other, assumption (1) fails.

Table 7 (Panel B) presents the parameters associated with assumption (2). In our context, this assumption requires $\lambda(1, h_1) - \lambda(0, h_1) = \lambda(1, 1) - \lambda(0, 1) = \lambda(1, 0) - \lambda(0, 0) = 0$ for $h \in \mathcal{H}_t$. This assumption implies that higher ability workers cannot enjoy additional returns to training across different time periods and training histories. We conduct the same χ^2 test and we find the three parameters are statistically different from each other.

As we have shown in this sub-section, the first-differences estimator fails to meet any of the standard assumptions required to identify the Average Treatment Effect of job training on wages. As first-differences and fixed-effect estimators are equivalent for $T = 2$ models, these estimators are not able to recover the ATE, which is the parameter of interest in this case. As a result, we argue that empirical strategies which rely on individual fixed effects will not recover the relevant parameters, and should thus be used with extreme caution in similar contexts to the one presented in this paper. In the next sub-section, we present the results of our structural model, which allows us to correctly recover the ATE parameter, as well as different dynamic treatment effect parameters.

5.2 Static Treatment Effects

Using our model estimates, we explore effect of training along different nodes. We first examine the effect of training in the first year on second-year earnings ($Y_{it}(1) - Y_{it}(0)$, for $t = 1$), and then analyze the effect of training in the second year on third-year earnings, conditional on the initial decision ($Y_{it}(1, h_t) - Y_{it}(0, h_t)$, $h_t \in \mathcal{H}_t$, for $t = 2$).[18] At the population level, we can define a variety of average treatment effects. To illustrate the benefits of our methodology, we focus on average treatment effects (ATE) and treatment on the treated effects (TT). Let $E[.]$ denote expected value taken with respect to the distribution of $(\mathbf{X}, \theta, \varepsilon)$, where ε is the collection of idiosyncratic shocks

[18] In a static analysis, we concentrate the analysis in a particular node, conditional on a specific training history. In contrast, the dynamic analysis compares earnings across years and potentially across different training branches (for example, we compare being trained twice versus not being trained at all).

determining outcomes and choices $\varepsilon \equiv (\varepsilon_j^I, \varepsilon_s^Y)$. We define ATEs as follows:[19]

$$ATE_1 \equiv E[Y_{i1}(1) - Y_{i1}(0)], \tag{14}$$

$$ATE_2(h) \equiv E[Y_{i2}(1, h_2) - Y_{i1}(0, h_2) \mid D_{i1} = h_2] \qquad \text{with } h_2 \in \mathcal{H}_2 \equiv \{0, 1\}. \tag{15}$$

Furthermore, reduced form estimation strategies do not allow us to estimate counter-factual outcomes for individuals in the sample, rendering us unable to estimate parameters of interest such the treatment on the treated (TT) parameter. As a result, we take advantage of our structural model to estimate TT along similar dimensions to those specified above. Treatment on the treated parameters (TT) equal:

$$TT_1 \equiv E[Y_{i1}(1) - Y_{i1}(0) \mid D_{i1} = 1], \tag{16}$$

$$TT_2(h) \equiv E[Y_{i2}(1, h_2) - Y_{i2}(0, h_2) \mid D_{i2} = 1, D_{i1} = h_2] \qquad \text{with } h_2 \in \mathcal{H}_2 \equiv \{0, 1\}. \tag{17}$$

We present static treatment effects in Table 8 (equations 14-17). The Table shows ATE and TT estimates (in rows) across different quantiles of unobserved ability θ (in columns).[20] The first column depicts the average effects for different quantiles of θ. For all individuals (ATE_1), we find a positive return to training on earnings associated with training in the first period. For individuals choosing training in $t = 1$ (TT_1), we also positive and statistically significant return to training. Moreover, we find larger effects for workers who had not chosen to be trained in the first period, but chose to do so in the second period. In fact, we find the largest return for lower ability workers, resembling the findings from the first training node. On the other hand, the returns to second-period training for workers who had already been trained is less than one percent. Nonetheless, we find significant heterogeneity in these returns: as we found above, there are large and positive returns for workers in the lowest ability quartile, but significant, and negative returns for highest ability workers. While the negative returns to training for high ability workers in the second training node may be surprising, this result can be understood within the framework of our dynamic model. Taking up a negative return can be explained by two reasons. First, the worker correctly predicts

[19] In our analysis, we use similar ATE parameters averaging across $\mathbf{X}$ and ε, and thus conditioning the ATE on θ.

[20] We show standard errors of our estimators in parenthesis. These standard errors are obtained from our simulated sample. The standard errors capture the variation of ATE and TT for a given structural parameters as well as parameters' uncertainty coming from the estimated posteriors.

a negative return, but chooses to take it anyways. In this case, our model suggests the existence of non-pecuniary benefits of training; in terms of our framework, the value of training, $I_it(h_t)$, is a weighted average of earnings and psychic costs (or benefits) associated with training. A second option is that individuals have systematic biases with respect to predicting the future economic returns to training. Both channels can be captured in our econometric model (in a reduced-form way).

Across all training nodes, we highlight heterogeneous returns to training, thus showing the inherent limitations of the fixed-effects approach in this context. In fact, we take advantage of our estimated model parameters and find that the first difference estimator in equation (9) would deliver an average treatment effect of 1.5 percent, which, as discussed above, is constant across all training nodes and histories. Given the results presented in this section, this treatment parameter fails to capture the heterogeneity in returns for workers' training choices in the first and second nodes.

The static returns presented above examine the returns to training in the year following training, but it is possible to instead focus the analysis down to a pairwise comparison. In this context, as workers are able to choose among four training histories, there are six different pairwise comparisons of earnings after the second training choice. As a result, we can compare the expected returns to training for workers who are trained in both years against a counterfactual training history of training in the first year and no training in the second, but also against a history of no training in the first followed by positive selection in the second year. In Table 9, we present the average treatment effect and the treatment on the treated parameters across the six counterfactual histories. The treatment on the treated parameter for training in the second year for first year trainees does not imply additional returns from subsequent training. Nonetheless, the earnings of the always trained group are still higher than the earnings of those in the never trained group. Interestingly, we also find that the timing of training matters, as workers who do not participate in the first period, but choose to do so in the second year outearn workers who train in the initial period, but fail to do so in their second year in the labor force. This result indicates a potential depreciation in the skills acquired through off-the-job training programs.

In sum, the estimated static treatment effects reveal heterogeneity across different decisions margins, labor training histories, and latent abilities. These findings bring support to a model of differential gains from training instead of the constant-effect framework of fixed-effect estimators.

5.3 Dynamic Treatment Effects

In this section, we extend our analysis by defining and estimating dynamic treatment effects. Moreover, we decompose dynamic treatment effects into direct effects and continuation values.[21] Hence, this section presents a clearer view on the real gains from training, who benefits from it, and the mechanisms behind the effects.

First, we define what we understand as dynamic treatment effects. In this case, we are interested in identifying the difference in outcomes—both in the short- and long-term—between participating in training and not. In terms of our model, we seek to identify the difference in the net present value of earnings associated to these two alternatives. While in this paper we focus on the effect of two training events, our framework can be extended to evaluate the impact of training decisions on life-time earnings. Formally, let $\widetilde{Y}_1(j)$ be the net present value associated with choosing training option j at period $t = 1$ ($D_1 = j$). This object is given by

$$\widetilde{Y}_{i1}(j) = Y_{i1}(j) + \rho \left(D_{i2}(j) Y_{i2}(1, j) + (1 - D_{i2}(j)) Y_{i2}(0, j) \right), \qquad j \in \mathcal{H}_2 \equiv \{0, 1\}$$

where ρ is the individual's discount factor. The individual-level dynamic treatment effect equals $\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0)$.

To better understand what a dynamic treatment effects comprises, let us express $\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0)$ as follows:

$$\begin{aligned} \widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0) = & \underbrace{(Y_{i1}(1) - Y_{i1}(0))}_{\text{Direct effect (short-term)}} + \underbrace{\rho [Y_{i2}(0, 1) - Y_{i2}(0, 0)]}_{\text{Direct effect (medium-term)}} + \\ & \underbrace{\rho \left[D_{i1}(1)(Y_{i2}(1, 1) - Y_{i2}(0, 1)) - D_{i1}(0)(Y_{i2}(1, 0) - Y_{i2}(0, 0)) \right]}_{\text{Continuation value}}. \end{aligned} \tag{18}$$

The first two terms of the right-hand side show the direct effect of training at $t = 1$ (properly discounted). The first of these terms represents the training return in earnings at following first period training (that is, in the short term), while the second the return to training on earnings at

[21] Heckman and Navarro (2007) and Heckman et al. (2016) perform a similar decomposition. We compare our results with that of Heckman et al. (2016).

after the completion of $t21$ (two years after training). The third term captures the continuation value of training. It represents the additional gain (if any) of training in $t = 2$ from training in $t = 1$. The continuation value captures potential dynamic complementarities: training in one period boosts the returns from training in future ones. The production function of training exhibit dynamic complementarities if the return from training in a second period is higher if the individual was trained in the first period $((Y_{i2}(1,1) > Y_{i2}(0,1)) - (Y_{i2}(1,0) - Y_{i2}(0,0)))$. Thus, we can test for dynamic complementarities by estimating ex-post continuation values as in Heckman et al. (2016).

Compared to a static approach, there are two additional components that a dynamic treatment captures: the medium-term effect and the continuation value. In a reduced-form approach, both of these terms are neglected or under-identified. In a fixed-effects set-up, continuation values are assumed away. In the causal-effect literature of RCTs, if longer-periods are ignored, then medium- and long-term effects are not taking into consideration.[22]

Moreover, the nature of these experiments does not allow for decomposing average effects into the three components of equation (18). Within our framework we can instead understand the mechanisms driving the dynamic effects of training. In what follows, we estimate versions of the average dynamic treatment effect (DATE):

$$DATE \equiv E[\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0)], \tag{19}$$

$$DTT \equiv E[\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0) \mid D_{i1} = 1], \tag{20}$$

$$DTUT \equiv E[\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0) \mid D_{i1} = 0], \tag{21}$$

and estimate their sub-components as well. Table 10 presents dynamic treatment effects (ATE, TT, TUT) and their decomposition into short- and medium-term direct effects and continuation values.[23] For an average worker, the dynamic returns to training are significant: the estimated DATE (equation 19) equals 2.91%, explained by a combination of short-term direct effects (period-two earnings), as well as medium term earnings. The negative continuation value indicates that

[22] Attanasio et al. (2011) evaluates the impact of the training program *Jóvenes en Acción* 13 and 15 months after the program ended. Card et al. (2011) looks at the effect of the program *Juventud y Empleo* in the Dominican Republic on outcomes 10 to 14 months after graduation. More recently, Alzúa et al. (2014) analyze the effect Argentina's *entra21* three years after it ended, and Attanasio et al. (2015) and Ibarrarán et al. (2015) have gathered longer-term outcomes from *Jóvenes en Acción* and *Juventud y Empleo*, respectively.

[23] We assume a discount factor $\rho = 1/(1.05)$.

training does not exhibit dynamic complementarities, which fits in with our finding that workers do not experience additional returns from second year training had they already been trained. We note there are no significant differences in the dynamic returns to training across workers who were in fact trained in the first year and those who were not. This result is driven by a combination of higher short- and medium-term returns for trained workers, alongside with lower continuation values for program participants.

Figure 6 depicts a decomposition of the DATE, DTT, and DTUT across unobserved ability deciles. We see that direct effects and continuation values differ by latent ability. For all types workers (treated or not), the medium-term effects are positive and increasing in ability, where as the short-term effects are large for low-ability workers, in the range of 4%, and decrease in ability. Moreover, continuation values decrease with ability and become negative for workers in the highest ability decile, consistent with the results presented in Table 10.

Overall, our results shed light on the production of human capital of training courses and reveal stark contrasts with the educational production function of other educational investments. Training does not have dynamic complementarities. In fact, training in the first period reduces the earnings return to training in a second-period. This result stands in contrast with what has been estimated for earlier educational instances. For example, Heckman et al. (2016) finds that the bulk of the return to high school graduation (around 70%) comes from continuation value. Moreover, this effect is stronger for high-ability students. We find just the opposite: continuation values are negative, especially for high-ability workers.

These patterns might be explained by the idiosyncrasies of Chile's training system. Following the literature, in this paper we view "training" under the lens of a unique production function. In practice, however, there are multiple types of training courses in Chile. Hence, one mechanism that could explain why low-ability workers gain the most from training is because *Franquicia Tributaria* tends to offer courses teaching basic skills, such as word processing, which may be complementary to these workers' underlying skill vector. Thus, the lack of quality and relevance might explain our findings. We consider our results a first step towards understanding the complex dynamic of the returns to training.

6 Conclusion

Given the ongoing changes in skill requirements across occupations, firms often find it necessary to train their workers on repeated occasions, especially for new entrants to the labor force. Nevertheless, estimating the impact of such training is often difficult due to limited data availability of firm-sponsored training events. In this paper, we have taken advantage of a large government-subsidized program to present the first estimates of repeated participation in off-the-job training for first-time labor market entrants. Our conceptual framework has allowed us to highlight the underlying assumptions required for fixed-effects estimators—a commonly-used strategy used in the training literature—to identify the average treatment effect of training. An extensive literature in labor economics has previously highlighted the importance of accounting for self-selection into job training participation, and various papers have controlled for selection on unobserved ability by estimating fixed effect models. In a dynamic model of training, we show that these estimators are only able to recover treatment effects under strong assumptions. These assumptions necessarily ignore the possibility of dynamic complementarities in the return to training.

To capture static and dynamic treatment effects, we introduce a dynamic econometric model of training participation and earnings. First, we test the required assumptions in a fixed effects strategy and document that these assumptions are not met in the data, thus showing that fixed-effects estimators are unable to estimate relevant treatment effect parameters. Our empirical results show that there is self-selection in all decision nodes, as high ability individuals are more likely to participate in *Franquicia Tributaria.* We find that our static treatment effect parameters vary depending on the relevant treatment node and by workers' ability. Our dynamic treatment effect parameters suggest positive returns to job training participation in the first year in the labor force, in the range of 3 percent. Nonetheless, we do not find that there are dynamic complementarities in the return to training, as first-year participants do not enjoy additional returns from their second training event. These results can be reconciled in a context where high ability workers are more likely to participate in training, but since the courses offered in the *Franquicia Tributaria* tend to be low-quality, they may enjoy decreasing returns from additional program participation. In this paper, we have highlighted the importance of estimating a number of different treatment parameters, as these vary across training nodes and worker ability. Finally, the estimated returns to training also vary depending on whether the estimated parameters are static or dynamic in nature, an important contribution of our paper.

References

Abbring, J. and G. van den Berg (2003). The non-parametric identification of treatment effects in duration models. *Econometrica 45*, 1491–1517.

Acemoglu, D. and D. Autor (2011). Chapter 12 Skills, Tasks and Technologies: Implications for Employment and Earnings. In D. Card and O. Ashenfelter (Eds.), *Handbook of Labor Economics*, Volume 4, pp. 1043–1171.

Acemoglu, D. and J.-S. Pischke (1999). Beyond becker: training in imperfect labour markets. *The Economic Journal 109*(453), 112–142.

Agostinelli, F. and M. Wiswall (2016a). Estimating the Technology of Children's Skill Formation. NBER Working Paper No. 22442.

Agostinelli, F. and M. Wiswall (2016b). Identification of Dynamic Latent Factor Models: The Implication of Re-Normalization in a Model of Child Development. NBER Working Paper No. 22441.

Alzúa, M. L., G. Cruces, D. D. T. Nro, and C. Lopez (2014). Youth Training Programs Beyond Employment. Experimental Evidence from Argentina. Documentos de Trabajo No. 177, CEDLAS.

Arellano, M. and B. Honore (2001). Panel Data Models: Some Recent Developments. In J. Heckman and E. Lazear (Eds.), *Handbook of Econometrics*.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics 60*(1), 47–57.

Ashenfelter, O. and D. Card (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistic 67*(4), 648–660.

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2015). Estimating the Production Function For Human Capital: Results From a Randomized Control Trail in Colombia. NBER Working Paper No. 20965.

Attanasio, O., A. Guarín, C. Medina, and C. Meghir (2015). Long Term Impacts of Vouchers for Vocational Training: Experimental Evidence for Colombia. NBER Working Paper 21390.

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*(3), 188–220.

Autor, D. H., F. Levy, and R. J. Murnane (2003, nov). *The Quarterly Journal of Economics 118*(4), 1279–1333.

Ba, B., J. C. Ham, R. J. Lalonde, and X. Li (2017). Estimating (easily interpreted) dynamic training effects from experimental data. *Journal of Labor Economics 35*(S1).

Booth, A. L. (1993). Private sector training and graduate earnings. *The Review of Economics and Statistics 75*, 164–170.

Buhrmann, J. R. (2017). Good enough jobs: Skill mismatch and two-sided heterogeneity in frictional labor markets.

Card, D. (1999). Chapter 30 The Causal Effect of Education on Earnings. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3, pp. 1801–1863.

Card, D., P. Ibarrarán, F. Regalia, D. Rosas-Shady, and Y. Soares (2011). The Labor Market Impacts of Youth Training in the Dominican Republic. *Journal of Labor Economics 29*(2), 267–300.

Card, D., J. Kluve, and A. Weber (2010, nov). Active Labour Market Policy Evaluations: A Meta-Analysis. *The Economic Journal 120*(548), F452–F477.

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003, may). 2001 Lawrence R. Klein Lecture Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice*. *International Economic Review 44*(2), 361–422.

Fitzenberger, B., M. Furdas, and C. Sajons (2016). End-of-year spending and the long-run employment effects of training programs for the unemployed. *IZA DP No. 10441*.

Frazis, H. and M. A. Loewenstein (2007). *On-the-Job Training.* Now Publishers, Inc.

Fredriksson, P. and P. Johansson (2008). Dynamic treatment assignment: The consequences for evaluations using observational data. *Journal of Business Economic Statistics 26*(4).

Guvenen, F., B. Kuruscu, S. Tanaka, and D. Wiczer (2015). Multidimensional skill mismatch. *Federal Reserve Bank of Minneapolis. Research Department Working Paper 729*.

Hansen, K. T., J. J. Heckman, and K. J. Mullen (2004). The effect of schooling and ability on achievement test scores. *Journal of Econometrics 121*(1), 39–98.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica 66*(5), 1017.

Heckman, J., R. Pinto, and P. Savelyev (2013, oct). Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review 103*(6), 2052–2086.

Heckman, J. J., J. E. Humphries, and G. Veramendi (2016). Dynamic treatment effects. *Journal of Econometrics 191*(2), 276–292.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999). The Economics and Econometrics of Active Labor Market Programs. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labour Economics: Volume 4B*, pp. 1865–2097. Elsevier Masson SAS.

Heckman, J. J. and S. Navarro (2007). Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics 136*(2), 341–396.

Heckman, J. J., J. Stixrud, and S. Urzua (2006, jul). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics 24*(3), 411–482.

Ibarrarán, P., J. Kluve, L. Ripani, and D. R. Shady (2015). Experimental Evidence on the Long-Term Impacts of a Youth Training Program. IZA Discussion Paper No. 9136.

Ingram, B. F. and G. R. Neumann (2006). The returns to skill. *Labour Economics 13*(1), 35–59.

Keane, M. P., P. E. Todd, and K. I. Wolpin (2011). Chapter 4 The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 4, pp. 331–461.

Lengermann, P. (1999). How long do the benefits of training last? evidence of long term effects across current and previous employers. *Research in Labor Economics 19*.

Lynch, L. M. (1992). Private-sector training and the earnings of young workers. *American Economic Review 82*, 299–312.

Magnac, T. and D. Thesmar (2002). Identifying Dynamic Discrete Decision Processes. *Econometrica 70*(2), 801–816.

Mincer, J. (1974). Schooling, experience, and earnings. human behavior & social institutions no. 2.

Rust, J. (1994). Structural Estimation of Markov Decision Processes. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume IV, Chapter 51, pp. 3139–3139.

Saltiel, F., M. Sarzosa, and S. Urzua (2018). Cognitive and Socio-emotional Ability. In G. Jones, J. Johnes, T. Agasisti, and L. Lopez-Torres (Eds.), *Handbook of Contemporary Education Economics*. Edwar Elgar Publishing.

Sanders, C. and C. Taber (2012, sep). Life-Cycle Wage Growth and Heterogeneous Human Capital. *Annual Review of Economics 4*(1), 399–425.

SpitzOener, A. (2006, apr). Technical Change, Job Tasks, and Rising Educational Demands: Looking outside the Wage Structure. *Journal of Labor Economics 24*(2), 235–270.

Veum, J. (1997). Training and job mobility among young workers in the united states. *Journal of Population Economics 10*, 219–233.

Appendix A: First Differences Estimator

In this section, we show how the FD estimator converges in probability to an average of $E[\Delta Y_{it}|\Delta D_{it} = -1]$ and $E[\Delta Y_{it}|\Delta D_{it} = -1]$:

$$\delta^{FD} \equiv 1/2 \times [E[\Delta Y_{it}|\Delta D_{it} = 1] - E[\Delta Y_{it}|\Delta D_{it} = -1]]$$

The FD estimator is:

$$\begin{aligned}
\delta^{FD} &\equiv \frac{Cov(\Delta Y_{it}, \Delta D_{it})}{Var(\Delta D_{it})} \\
&\equiv \frac{E[\Delta Y_{it} \times \Delta D_{it}] - E[\Delta Y_{it}] \times E[\Delta D_{it}]}{E[\Delta D_{it}] - (E[\Delta D_{it}])^2} \\
&\equiv \frac{E[\Delta Y_{it}|\Delta D_{it} = 1] \Pr(\Delta D_{it} = 1) [1 - (\Pr(\Delta D_{it} = 1) - \Pr(\Delta D_{it} = -1))]}{(\Pr(\Delta D_{it} = 1) + \Pr(\Delta D_{it} = -1)) - (\Pr(\Delta D_{it} = 1) - \Pr(\Delta D_{it} = -1))^2} \\
&\quad - \\
&\quad \frac{E[\Delta Y_{it}|\Delta D_{it} = -1] \Pr(\Delta D_{it} = -1) [1 + (\Pr(\Delta D_{it} = 1) - \Pr(\Delta D_{it} = -1))]}{(\Pr(\Delta D_{it} = 1) + \Pr(\Delta D_{it} = -1)) - (\Pr(\Delta D_{it} = 1) - \Pr(\Delta D_{it} = -1))^2}
\end{aligned}$$

After some algebraic manipulation and since $\Pr(\Delta D_{it} = 1) + \Pr(\Delta D_{it} = -1) = 1$, we get:

$$\begin{aligned}
\delta^{FD} &\equiv \frac{E[\Delta Y_{it}|\Delta D_{it} = 1] \Pr(\Delta D_{it} = 1) [1 - (\Pr(\Delta D_{it} = 1) - \Pr(\Delta D_{it} = -1))]}{1 - 4\Pr(\Delta D_{it} = 1)^2 + 4\Pr(\Delta D_{it} = 1) - 1} \\
&\quad - \\
&\quad \frac{E[\Delta Y_{it}|\Delta D_{it} = -1] \Pr(\Delta D_{it} = -1) [1 + (\Pr(\Delta D_{it} = 1) - \Pr(\Delta D_{it} = -1))]}{1 - 4\Pr(\Delta D_{it} = 1)^2 + 4\Pr(\Delta D_{it} = 1) - 1} \\
&\equiv \frac{2 \times E[\Delta Y_{it}|\Delta D_{it} = 1] \times \Pr(\Delta D_{it} = 1) \times \Pr(\Delta D_{it} = -1)}{4\Pr(\Delta D_{it} = 1) \times (1 - \Pr(\Delta D_{it} = -1))} \\
&\quad - \\
&\quad \frac{2 \times E[\Delta Y_{it}|\Delta D_{it} = -1] \times \Pr(\Delta D_{it} = 1) \times \Pr(\Delta D_{it} = -1)}{4\Pr(\Delta D_{it} = 1) \times (1 - \Pr(\Delta D_{it} = -1))} \\
&\equiv \frac{E[\Delta Y_{it}|\Delta D_{it} = -1] - E[\Delta Y_{it}|\Delta D_{it} = -1]}{2}
\end{aligned}$$

Appendix B: FD Estimator and Treatment Effects

In this section, we prove Theorem 1. Let h and h' denote elements of $\mathcal{H}_t$ and $\mathcal{H}_{t-1}$. We can express the FD estimator as

$$\begin{aligned}\delta^{FD} &= 1/2 \times E\left[\sum_{h\in\mathcal{H}_t} H_{it}(h)Y_{it}(1,h) - \sum_{h'\in\mathcal{H}_{t-1}} H_{it-1}(h')Y_{it-1}(0,h')\right] \\ &- 1/2 \times E\left[\sum_{h\in\mathcal{H}_t} H_{it}(h)Y_{it}(0,h) - \sum_{h'\in\mathcal{H}_{t-1}} H_{it-1}(h')Y_{it-1}(1,h')\right]\end{aligned}$$

Given our assumption about counterfactual outcomes (equation 2), the equation above reduces to:

$$\begin{aligned}\delta^{FD} &= 1/2 \times E\left[\sum_{h\in\mathcal{H}_t} H_{it}(h)(\mu^Y(1,h) + \lambda^Y(1,h)\theta_i) - \sum_{h'\in\mathcal{H}_{t-1}} H_{it-1}(h')(\mu^Y(0,h') + \lambda^Y(0,h')\theta_i)\right] \\ &- 1/2 \times E\left[\sum_{h\in\mathcal{H}_t} H_{it}(h)(\mu^Y(0,h) + \lambda^Y(0,h)\theta_i) - \sum_{h'\in\mathcal{H}_{t-1}} H_{it-1}(h')(\mu^Y(1,h') + \lambda^Y(1,h')\theta_i)\right],\end{aligned}$$

and collecting terms, we have

$$\begin{aligned}\delta^{FD} &= 1/2 \times E\left[\sum_{h\in\mathcal{H}_t} H_{it}(h)(\mu^Y(1,h) - \mu^Y(0,h)) + \sum_{h\in\mathcal{H}_t} H_{it}(h)(\lambda^Y(1,h) - \lambda^Y(0,h))\theta_i\right] \\ &+ 1/2 \times E\left[\sum_{h'\in\mathcal{H}_{t-1}} H_{it-1}(h')(\mu^Y(1,h') - \mu^Y(0,h')) + \sum_{h'\in\mathcal{H}_{t-1}} H_{it-1}(h')(\lambda^Y(1,h') - \lambda^Y(0,h'))\theta_i\right].\end{aligned}$$

Reducing the expression above by applying the expected value operator cannot yield ATE, because of two fundamental reasons. First, $H_t(h)$ is, in general, not independent of θ_i, since agents may sort into training at different periods based on their knowledge of θ_i. Second, even if $H_t(h)$ and θ_i were independent, the resulting weighted averages of treatment effects of t and $t-1$ may not necessarily have to be the same. Under assumptions 1 and 2, the second term in each square bracket collapses to 0 and the first term to a constant π_1. We have then

$$\delta^{FD} = 1/2 \times \pi_1 + 1/2 \times \pi_1 = \pi_1$$

As a result, assumptions 1 and 2 imply that the ATE equals precisely π_1 across all training nodes and training histories (see equation 4).

Table 1: Variables in the Implementation of the Model

Variables	Wage Equation	Training Probit	Measurement System
Constant	Yes	Yes	Yes
Gender	Yes	Yes	Yes
Age at Test			Yes
Age in Year t	Yes	Yes	
Cognitive Factor	Yes	Yes	Yes
Household Size			Yes
Mother's Education			Yes
Father's Education			Yes
Mother's Employment			Yes
Father's Employment			Yes
Average Training Hours at Firm		Yes	

Note: We show the variables used in our empirical model. In the measurement system, we use math, language and high school GPA along with a gender dummy and the age at the time of the test. For the training participation model, we include the worker's gender, age in year t, the cognitive factor along with the instrumental variable which captures the average hours of *Franquicia Tributaria* used by the employing firm at the time of training decisions.

Table 2: Summary Statistics for Full Sample

Variable	**Mean**	**(Std. Dev.)**
Male	0.41	(0.49)
Age at Graduation	17.76	(0.63)
Math PSU (Normalized)	0.18	(1.05)
Verbal PSU (Normalized)	0.17	(1.03)
High School GPA (Normalized)	0.12	(1.02)
Monthly Salary after First Year (USD)	656.02	(407.28)
Monthly Salary after Second Year (USD)	749.55	(468.56)
N	5000	

Note: Table 2 presents summary statistics for a sample of 10,000 workers who were employer in firms which were eligible for *Franquicia Tributaria* subsidies in their first two years in their labor force. As noted in the text, the salary variable measures the monthly average of earnings in the first quarter following the training period, but for notational simplicity, we refer to this variable as concurrent with the training decision. Note that both test score measures (Math and Verbal) and the high school GPA are normalized across the general population of test-takers to be (0,1).

Table 3: Summary Statistics by Training Node

	Tree = 0,0	Tree = 0,1	Tree = 1,0	Tree = 1,1
Male	0.41	0.43	0.41	0.38
Age at Graduation	17.76	17.73	17.75	17.77
Math PSU (Normalized)	0.10	0.29	0.18	0.45
Verbal PSU (Normalized)	0.10	0.30	0.15	0.40
High School GPA (Normalized)	0.09	0.20	0.09	0.22
Monthly Salary after First Year (USD)	592.07	727.89	666.54	878.41
Monthly Salary after Second Year (USD)	679.24	838.41	761.37	981.95
Observations	3041	731	619	609

Note: Table 3 presents summary statistics for a sample of 10,000 workers who were employer in firms which were eligible for *Franquicia Tributaria* subsidies in their first two years in their labor force divided by their training choices in these two years. The salary variable measures the monthly average of earnings in the first quarter following the training period, but for notational simplicity, we refer to this variable as concurrent with the training decision. Both test score measures (Math and Verbal) and the high school GPA are normalized across the general population of test-takers to be (0,1).

Table 4: Structural Model: Loadings on Measurement Tests

	Math PSU	Language PSU	High School GPA	Initial Salary
Constant	2.43	2.97	3.86	-2.69
	(0.38)	(0.37)	(0.39)	(0.15)
Male	0.15	-0.07	-0.36	0.05
	(0.03)	(0.03)	(0.03)	(0.03)
Age	-0.17	-0.19	-0.21	0.12
	(0.02)	(0.02)	(0.02)	(0.01)
Household Size	-0.02	-0.04	0.01	.
	(0.01)	(0.01)	(0.01)	(.)
Mother's Education	0.04	0.05	0.01	.
	(0.00)	(0.00)	(0.00)	(.)
Father's Education	0.04	0.04	0.00	.
	(0.00)	(0.00)	(0.00)	(.)
Father's Employment	-0.09	-0.13	-0.09	.
	(0.03)	(0.03)	(0.04)	(.)
Mother's Employment	-0.00	0.03	-0.08	.
	(0.03)	(0.03)	(0.03)	(.)
Cognitive Factor	0.46	0.34	0.35	1.00
	(0.02)	(0.02)	(0.02)	(0.00)
Precision	1.21	1.20	1.11	4.86
	(0.02)	(0.02)	(0.02)	(0.12)
Observations	1000	1000	1000	1000

Note: Table 4 displays the estimation results from the measurement system of test scores. The dependent variable is the normalized (0,1) test score. Male, father's and mother's employment are dummy variables, whereas age, household size, mother's and father' education are coded as continuous variables. Standard errors are in parentheses. The loading on cognitive factor in the Math PSU is normalized to 1.

Table 5: Goodness of Fit Tests: Average Log Wages by Year, Training Node

	Y_1	Y_2	$Y_2\|1,1$	$Y_2\|1,0$	$Y_2\|0,1$	$Y_2\|0,0$
A. Means						
Actual	12.71	12.83	12.98	12.86	12.92	12.78
Model	12.71	12.83	13.00	12.84	12.92	12.79
B. Standard Deviations						
Actual	0.57	0.59	0.61	0.60	0.60	0.58
Model	0.55	0.57	0.55	0.58	0.55	0.55
C. P-value of test	0.960	0.581	0.001	0.001	0.271	0.161

Note: In Table 5, we show the means of earnings by year and schooling choice from the data and the simulated sample. In each cell, we also show the p-value for a two-sample mean-comparison test, in which we fail to reject the equality of observed and simulated average monthly earnings.

Table 6: Reduced Form Results: Horse Race

	(1) OLS	(2) OLS	(3) First Differences	(4) FD: Interactions	(5) Individual+Firm FE
Training	0.249***	0.194***	0.0166*	0.0154*	0.0215**
	(0.000)	(0.000)	(0.028)	(0.048)	(0.002)
Math		0.141***		0	
		(0.000)		(.)	
Verbal		0.00862		0	
		(0.224)		(.)	
GPA		0.0609***		0	
		(0.000)		(.)	
Male		0.101***			
		(0.000)			
Age		0.0979*			
		(0.040)			
Age^2		0.0000368			
		(0.973)			
Training=1				0	
				(.)	
Training=1 X Math				-0.00739	
				(0.492)	
Training=1 X Verbal				0.00473	
				(0.670)	
Training=1 X GPA				0.0127	
				(0.145)	
Constant	12.71***	10.41***	12.77***	12.77***	
	(0.000)	(0.000)	(0.000)	(0.000)	
Observations	10000	10000	10000	10000	10000

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Testing Assumptions 1 and 2 for Validity of FD Estimators

(a) Assumption 1: $\mu^Y(1,h_t) - \mu^Y(0,h_t) = \pi_1$

		$\mu^Y(1,h_1) - \mu^Y(0,h_1)$	$\mu^Y(1,0) - \mu^Y(0,0)$	$\mu^Y(1,1) - \mu^Y(0,1)$
Coefficient		0.019	0.030	0.011
Standard Error		0.037	0.031	0.046
p-value of test	0.000			

(b) Assumption 2: $\lambda^Y(1,h_t) - \lambda^Y(0,h_t) = \pi_1$

		$\lambda^Y(1,h_1) - \lambda^Y(0,h_1)$	$\lambda^Y(1,9) - \lambda^Y(0,9)$	$\lambda^Y(1,1) - \lambda^Y(0,1)$
Coefficient		-0.015	-0.019	-0.040
Standard Error		0.001	0.011	0.015
p-value of test	0.000			

Note: In Table 7, we show the means of earnings by year and schooling choice from the data and the simulated sample. In each cell, we also show the p-value for a two-sample mean-comparison test, in which we fail to reject the equality of observed and simulated average monthly earnings.

Table 8: Static treatment Effects: Effects on log earnings (in percentage points)

	All	**Q1**	**Q2**	**Q3**	**Q4**
A. $t = 2$ (after first decision)					
ATE_1	1.95	3.53	2.45	1.53	0.28
	(0.060)	(0.12)	(0.12)	(0.12)	(0.12)
TT_1	2.01	3.17	2.62	2.09	0.62
	(0.000)	(0.27)	(0.26)	(0.25)	(0.23)
B. $t = 3$ (after second decision)					
$ATE_2(0)$	3.10	4.93	3.65	2.52	1.09
	(0.120)	(0.230)	(0.230)	(0.240)	(0.240)
$TT_2(0)$	3.20	6.02	4.05	2.84	0.63
	(0.260)	(0.569)	(0.525)	(0.513)	(0.498)
$ATE_2(1)$	0.77	4.92	2.69	-0.15	-2.97
	(0.200)	(0.430)	(0.410)	(0.390)	(0.370)
$TT_2(1)$	0.40	4.78	2.58	-0.10	-2.63
	(0.290)	(0.746)	(0.623)	(0.552)	(0.492)

Note: In Table 8, we estimate static Average Treatment Effects (ATE), and Treatment on the Treated (TT) parameters across three training nodes in the first two years of workers' participation in the formal labor sector. We additionally take advantage of the distribution of the estimated ability factor and divide the sample by ability quartiles to estimate heterogeneous average treatment effects as well as treatment on the treated parameters.

Table 9: Treatment Effects: Returns to Job Training

Comparison	ATE	TT
Y_{11} vs. Y_{10}	0.011	0.004
	(0.001)***	(0.003)***
Y_{11} vs. Y_{01}	-0.005	-0.003
	(0.001)***	(0.003)
Y_{11} vs. Y_{00}	0.026	0.027
	(0.001)***	(0.004)***
Y_{10} vs. Y_{01}	-0.016	-0.018
	(0.001)	(0.002)
Y_{10} vs. Y_{00}	0.016	0.017
	(0.001)*	(0.002)**
Y_{01} vs. Y_{00}	0.031	0.032
	(0.001)***	(0.001)***

Note: In Table 9, we compare different static treatment effect parameters following training in the second year using our simulated parameters from our model.

Table 10: Dynamic treatment Effects: Effects on log earnings (in percentage points)

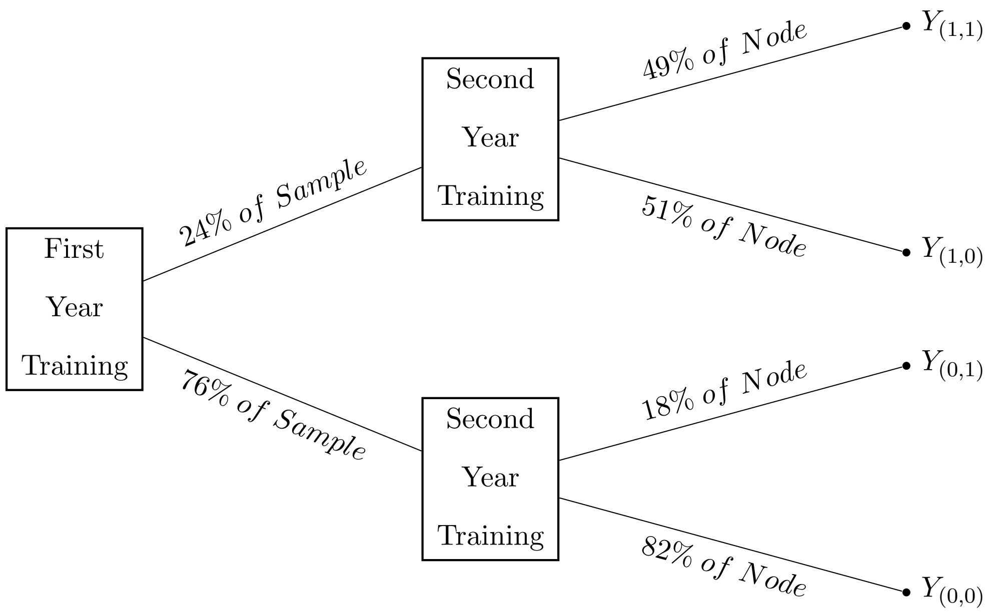
	DATE	**DTT**	**DTUT**
Direct effect (short-term)	1.95	2.01	1.92
	(0.06)	(0.12)	(0.07)
	[67%]	[66%]	[67%]
Direct effect (medium-term)	1.45	1.87	1.32
	(0.10)	(0.20)	(0.11)
	[50%]	[62%]	[46%]
Continuation value	-0.48	-0.83	-0.37
	(0.07)	(0.16)	(0.08)
	[-16%]	[-27%]	[-13%]
Total	2.91	3.04	2.87
	(0.11)	(0.23)	(0.13)

Note: We estimate Dynamic Average Treatment Effects (DATE), Dynamic Treatment on the Treated (DTT) and Dynamic Treatment on the Untreated (DTUT) of training in $t = 1$ on the present value of earnings. We estimate

$$\begin{aligned} DATE &\equiv E[\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0)], \\ DTT &\equiv E[\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0) \mid D_{i1} = 1], \\ DTUT &\equiv E[\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0) \mid D_{i1} = 0], \end{aligned}$$

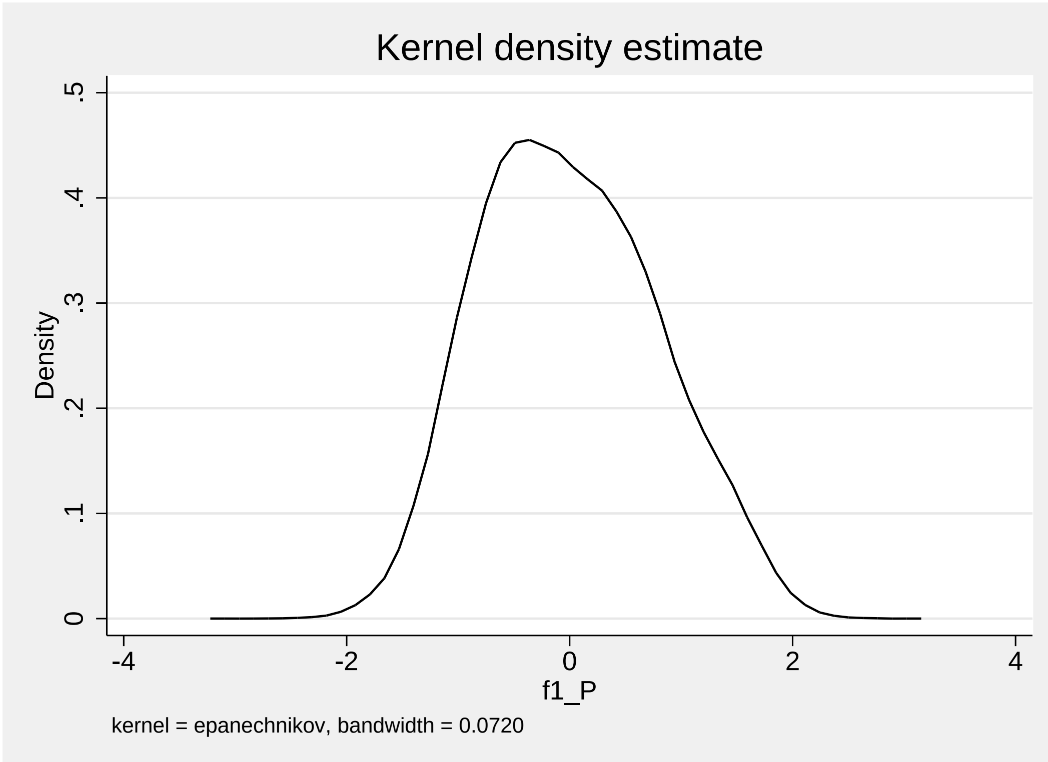
and decompose them into short-term direct effects $(\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0))$, medium-term direct effects $(\rho[Y_{i2}(0,1) - Y_{i2}(0,0)])$ and continuation value $(\rho\ E[D_{i1}(1)(Y_{i2}(1,1) - Y_{i2}(0,1)) - D_{i1}(0)(Y_{i2}(1,0) - Y_{i2}(0,0))])$.

Figure 1: Observed Training Choices



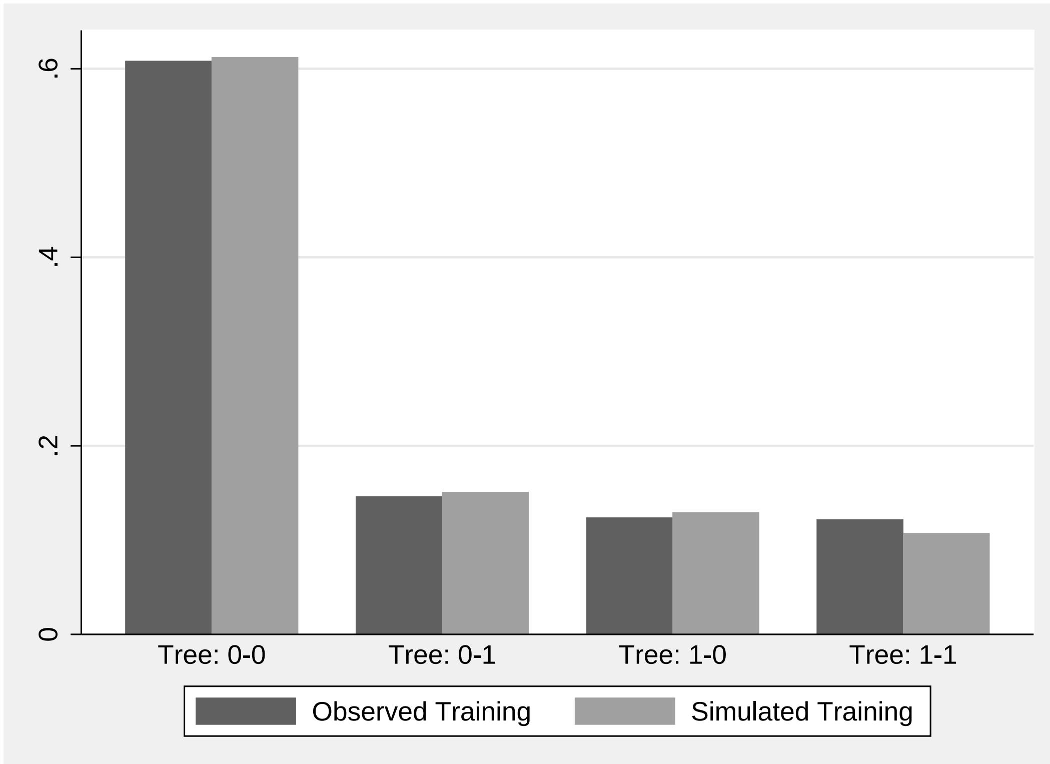
Note: Figure 1 presents the decision tree through which workers decide whether to participate in training in each of their first two years in the labor force. In each node, we include the observed share of workers in our sample who decide to either participate or not in training, eventually reaching one of four possible states after their first two years in the labor force.

Figure 2: Distribution of Factor



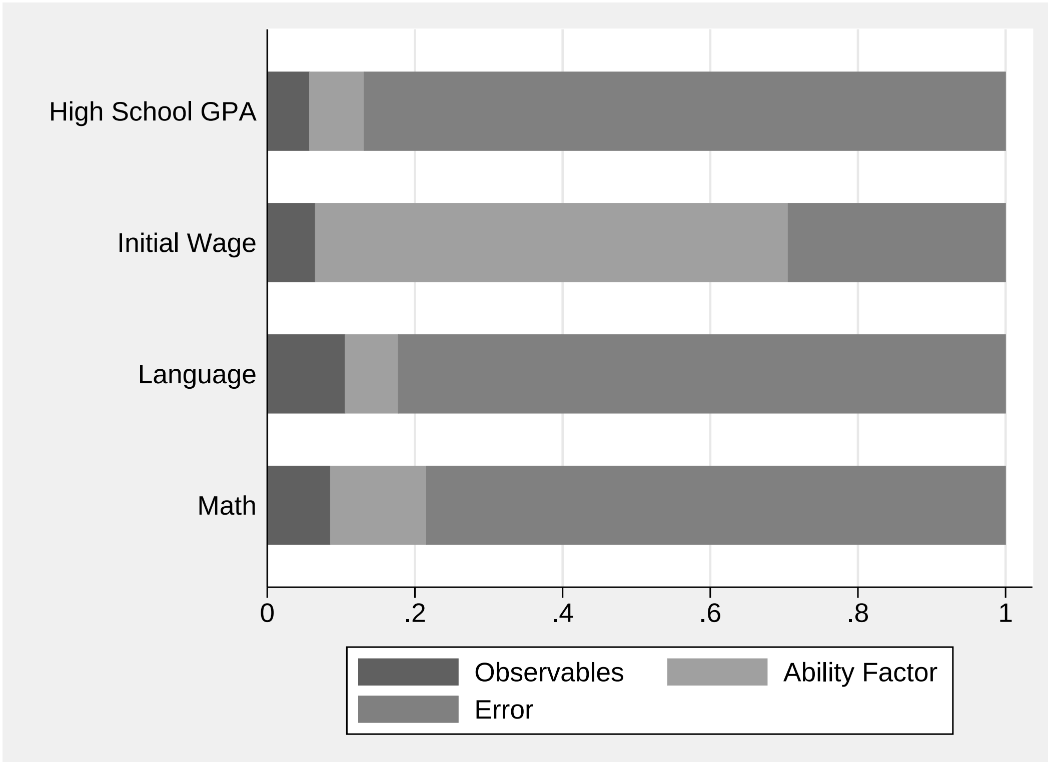
Note: In Figure 2, we show the estimated density of the individual-level unobserved ability. We obtain this density using the simulated sample from our estimated model. The sample size is 200,000. We approximate the distribution of the individual's unobserved ability factor by a three-component mixture of normal distributions.

Figure 3: Goodness of Fit: Training Decisions



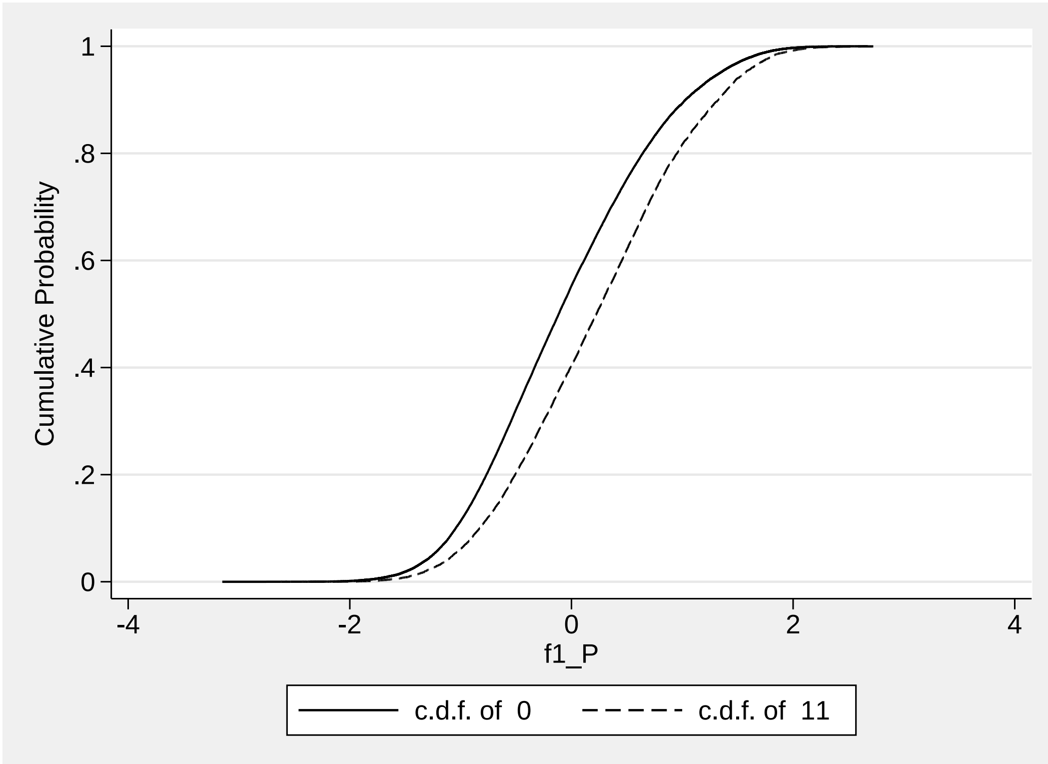
Note: In Figure 3, we compare the share of workers who followed each of the four possible training histories in their first two years in the labor force in the observed data against our simulated sample with 200,000 observations.

Figure 4: Variance Decomposition



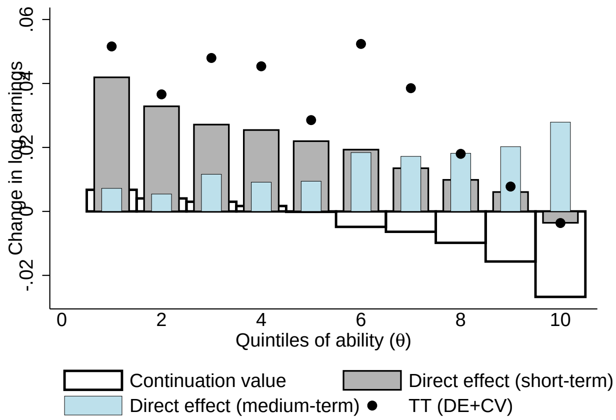
In Figure 4, we show the contribution of each variable to the variance of test scores using the simulated sample from our model. The row Observables indicates the share of the variance of the measurement variables explained by the observed variables: age at the time of test score, dummy variables for male and parent employment, and continuous variables in mother's and father's education, as well as household size. Ability Factor shows the proportion of the test score variance explained by the unobserved factor. Finally, the label Error term represents the share of each test score variance explained by the unobserved idiosyncratic error of the training participation model.

Figure 5: Distribution of Factor by Training History

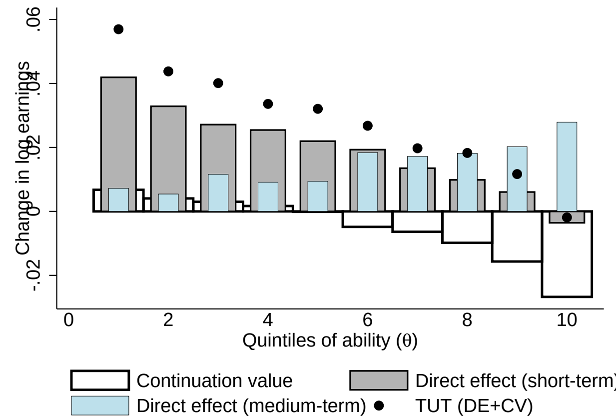


Note: Figure 5 shows the cumulative distribution of the unobserved factor for a specific training history at the end of the second year in the labor force. The results follow from our simulated sample using 200,000 observations.

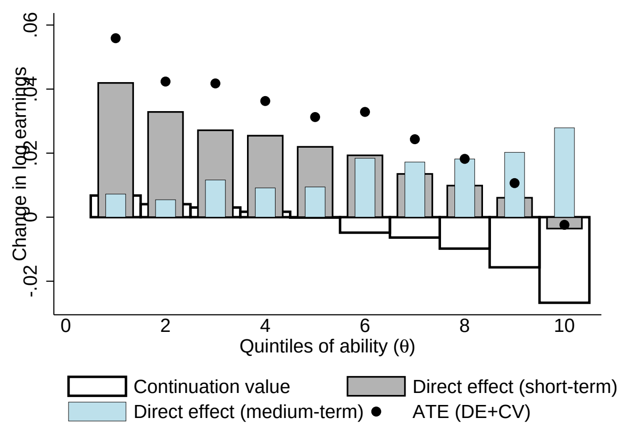
Figure 6: Dynamic treatment effects and latent ability



(a) DTT



(b) DTUT



(c) DATE

Note: We estimate Dynamic Average Treatment Effects (DATE), Dynamic Treatment on the Treated (DTT) and Dynamic Treatment on the Untreated (DTUT) of training in $t = 1$ on the present value of earnings. We estimate

$$\begin{aligned} DATE &\equiv E[\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0)], \\ DTT &\equiv E[\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0) \mid D_{i1} = 1], \\ DTUT &\equiv E[\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0) \mid D_{i1} = 0], \end{aligned}$$

and decompose them into short-term direct effects $(\widetilde{Y}_{i1}(1) - \widetilde{Y}_{i1}(0))$, medium-term direct effects $(\rho[Y_{i2}(0,1) - Y_{i2}(0,0)])$ and continuation value $(\rho\ E[D_{i1}(1)(Y_{i2}(1,1) - Y_{i2}(0,1)) - D_{i1}(0)(Y_{i2}(1,0) - Y_{i2}(0,0))])$.