A causal latent transition model with multivariate outcomes and unobserved heterogeneity: Application to human capital development

Francesco Bartolucci* Fulvia Pennoni† Giorgio Vittadini‡

15 December 2022

Abstract

In order to evaluate the effect of a policy or treatment with pre- and post-treatment outcomes, we propose an approach based on a transition model, which may be applied with multivariate outcomes and accounts for unobserved heterogeneity. This model is based on potential versions of discrete latent variables representing the individual characteristic of interest and may be cast in the hidden (latent) Markov literature for panel data. Therefore, it can be estimated by maximum likelihood in a relatively simple way. The approach extends the Difference-in-Difference method as it is possible to deal with multivariate outcomes. Moreover, causal effects may be expressed with respect to transition probabilities. The proposal is validated through a simulation study, and it is applied to evaluate educational programs administered to pupils in the 6th and 7th grades during their middle school period. These programs are carried out in an Italian region to improve non-cognitive skills. We study if they impact also on students' cognitive skills in Italian and Mathematics in the 8th grade, exploiting the pre-treatment test scores available in the 5th grade. The main conclusion is that the educational programs aimed to develop non-cognitive abilities help the best students to maintain their higher cognitive abilities over time.

KEYWORDS: causal inference, cognitive skills, hidden Markov models, human capital, non-cognitive skills

*Department of Economics, University of Perugia, Via A. Pascoli, 20, 06123 Perugia, email: francesco.bartolucci@unipg.it

†Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via degli Arciboldi, 8, 20126 Milano, email: fulvia.pennoni@unimib.it

‡Department of Statistics and Quantitative Methods and member of the Scientific Council of CRISP (Interuniversity Research Centre on Public Services), University of Milano-Bicocca, Via degli Arciboldi, 8, 20126 Milano, email: giorgio.vittadini@unimib.it

1 Introduction

In many applications, especially in education, the main focus is on the causal effect of a treatment or policy on a certain individual characteristic of interest, such as the ability in certain subjects. Even in absence of experimental data, a context in which this evaluation may be performed getting rid of different types of confounding factors is when pre- and post-treatment outcomes are available. In this context, it is natural to apply the Difference-in-Difference (DiD) method, which is also very popular in other fields such as economics (for a review, see Imbens and Wooldridge, 2009; Lechner, 2011; Lee, 2016).

Taking inspiration from the standard DiD method, we propose a novel causal inference approach based on potential versions of discrete latent variables that represent the individual characteristic of interest. It is based on a model that we name Causal Latent Transition (CLT) and that, in reduced form, is equivalent to a latent Markov (LM) model for panel data with initial and transition probabilities depending on individual covariates (Bartolucci et al., 2014).

The main features that characterize the CLT model are:

(*i*) *multivariate outcomes*: the model can be used in a multivariate setting where the same individual characteristic, represented by the latent variables, is measured by more response variables that may also have a different nature;

(*ii*) *unobserved heterogeneity*: the individuals are clustered in a finite number of homogenous subpopulations identified by the states of the latent variables that, by definition, are not directly observable. Specific causal effects are defined and estimated for each of these subpopulations.

To better understand the CLT model, it is useful to recall the principal characteristics of LM models for panel data (Bartolucci et al., 2013). These models have a structure closely related to that of hidden Markov models for time-series (MacDonald and Zucchini, 2016) as a sequence of discrete latent variables is assumed to exist for every individual. Each sequence follows a Markov chain of first order with a number of states that is left unspecified, thus providing more flexibility with respect to the corresponding models formulated on the basis of continuous latent processes (Bartolucci et al., 2022). In the

application motivating this paper these latent variables represent the individual characteristic or personal trait of interest, which is a certain type of cognitive ability. For every time occasion and conditionally on the corresponding latent variable, each response variable measuring an individual characteristic is conditionally independent of the response variables at different time occasions. As such, the CLT model may also be adopted with more than two occasions of observation, although in the application we use to illustrate the proposal only pre- and post-treatment outcomes are available.

Unlike the standard LM formulation, we adopt potential latent variables for the proposed CLT model to enhance causal interpretations. In particular, causal effects are expressed in terms of logits for the transition probabilities between states of these latent variables. However, it is also possible to express these effects in terms of differences between probabilities or directly as effects on the response variables in a flexible way. The idea of using potential versions of the latent variables in formulating an LM model has already been exploited in the model proposed by Bartolucci et al. (2016) that, in turn, extends the causal latent class model proposed by Lanza et al. (2013). In these approaches, model estimation is based on propensity score weights (Rosenbaum and Rubin, 1983; Rosenbaum, 2020). In contrast, in the current proposal, estimation of the causal effects is performed by directly including the covariates in the latent process so that certain types of unobserved confounding may be eliminated, as in the DiD approach. Moreover, the CLT model allows analyzing sequential stage developments from the estimated transition probabilities.

The model parameters are estimated by a rather standard Expectation-Maximization (EM) algorithm that makes use of suitable recursions (Baum et al., 1970; Dempster et al., 1977; Welch, 2003) so that the overall approach is relatively easy to apply even when extended to more than two time occasions. In particular, available statistical packages, such as `LMest` (Bartolucci et al., 2017) in the open source software R (R Core Team, 2022), may be directly used with minor adjustments; the code developed for the application is available at the GitHub repository: `https://github.com/penful/CausalLT`.

The proposed approach is validated by a simulation study and illustrated by an application aimed to analyze the effect of a certain treatment on the Human Capital (HC)

development, which comprises skills and expertises acquired through the investment in education and whose returns are identified by higher individual expected earnings (Becker, 1994). The HC has traditionally been defined in terms of Cognitive Skills (CSs), namely innate and acquired abilities and competencies usually associated with learning and problem solving tasks, such as reasoning, remembering, speaking, and understanding (see, among others, Heckman et al., 2014; OECD, 2015). However, researchers and practitioners in education have recently become more and more interested in measuring and studying Non-Cognitive Skills (NCSs) that, differently from the CSs, are defined as personality resources linked to motivation in learning, relational capabilities, emotional stability, and autonomy in pursuing personal objectives. NCSs potentially affect goal-directed efforts, healthy social relations, adequate judgement and decision-making; these skills can be improved by means of suitable educational programs (Heckman and Kautz, 2012; Heckman et al., 2014). A vast literature demonstrates that educational programs can increase the NCSs and that an increase in the NCSs produces a consistent improvement in the CSs. Therefore, in our application, we consider this hypothesis rising from the HC literature, and we address the following scientific question: "Do NCSs programs causally determine an improvement of the CSs?". To address this question, we rely on data coming from a study based on a sample of primary and middle class students of the Autonomous Province of Trento (named PAT) in Italy over three consecutive school years (from 2015 to 2018) in which students' cognitive abilities are measured at two occasions. During this period, the PAT implemented a plan based on educational activities tailored to reinforcing the NCSs of students. Data are referred to the schools that voluntarily agreed to this program so that we dispose of a sample involving treated and untreated PAT students. The effects of these programs are evaluated by considering Italian and Mathematics test scores derived from administrative surveys managed by the Italian National Institute for the Evaluation of the Educational System (INVALSI). Merging these data with those deriving from administrative surveys carried out by the PAT, we dispose of many covariates that can be suitably exploited.

The remainder of the paper is structured as follows. In Section 2, after a brief review of the LM model with covariates, we introduce the proposed CLT model, whose main

features are discussed in Section 3. In Section 4 we show the results of the simulation study, some details of which are reported in the Supplementary Information (SI) file. In Section 5 we introduce the application illustrating the NCSs and educational programs, and we describe the data. In Section 6 we report the empirical results of the CLT model and those obtained with the DiD method for the data at issue. Additional details and results related to the application are shown in the SI file. Finally, Section 7 provides main conclusions.

2 Causal latent transition model

In the following, after a brief review of the LM model with covariates in the structural component model, we describe the proposed CLT model, illustrating first its assumptions, its possible extensions, and finally the estimation method of the model parameters.

2.1 Preliminaries

In the context of a panel study and with reference to individual i, $i = 1, \ldots, n$, and occasion t, $t = 0, \ldots, T-1$, we observe a vector of r response variables $\boldsymbol{Y}_{it} = (Y_{i1t}, \ldots, Y_{irt})'$ that may be of different types. In the applicative context that will be illustrated in Section 5 these variables are continuous, but they may be categorical or discrete with an arbitrary number of levels. For every individual i we also consider a vector of time-varying covariates $\boldsymbol{X}_{it}$.

In order to model panel data having the structure described above, the LM approach (Bartolucci et al., 2013, 2014) relies on individual sequences of discrete latent variables that are collected in the vectors $\boldsymbol{H}_i = (H_{i0}, \ldots, H_{i,T-1})'$, $i = 1, \ldots, n$. Every latent variable H_{it} may assume a value from 1 to k; this amounts to define k latent states, or equivalently latent clusters or classes, with individuals in the same state having the same behavior. The latent variables affect the distribution of the corresponding vector of response variables so that each $\boldsymbol{Y}_{it}$ is conditionally independent of the other response vectors $\boldsymbol{Y}_{is}$, $s \neq t$, given H_{it}. The conditional distribution of $\boldsymbol{Y}_{it}$ given H_{it} may be of any type as in a finite mixture model (McLachlan and Peel, 2000). Analogously to our

proposal, mixture models assume that the sample is generated by different subpopulations or clusters thus extending the model-based clustering methods also known as unsupervised learning (Früuhwirth-Schnatter et al., 2019). When the response variables are continuous, it is natural to rely on the multivariate Gaussian distribution with mean depending on the latent state and common variance-covariate matrix (Bouveyron et al., 2002), that is,

$$\boldsymbol{Y}_{it}|H_{it}=h\sim N_r(\boldsymbol{\mu}_h,\boldsymbol{\Sigma}),\quad h=1,\ldots,k,\ i=1,\ldots,n,\ t=0,\ldots,T-1, \tag{1}$$

where latent state h is a realization of H_{it}. In certain formulations with categorical response variables, it is also assumed that the random variables of each vector $\boldsymbol{Y}_{it}$ are conditionally independent given H_{it}.

Every sequence $\boldsymbol{H}_i$ follows a first-order Markov chain with initial and transition probabilities depending on the covariates. In particular, we adopt the following multinomial logit parametrization for the initial probabilities:

$$\log\frac{p(H_{i0}=h|\boldsymbol{X}_{it}=\boldsymbol{x})}{p(H_{i0}=1|\boldsymbol{X}_{it}=\boldsymbol{x})}=\boldsymbol{x}'\boldsymbol{\beta}_h,\quad h=2,\ldots,k. \tag{2}$$

For the transition between states, the following multinomial logit parametrization is assumed for $t=1,\ldots,T-1$:

$$\log\frac{p(H_{it}=h|H_{i,t-1}=\bar{h},\boldsymbol{X}_{it}=\boldsymbol{x})}{p(H_{it}=\bar{h}|H_{i,t-1}=\bar{h},\boldsymbol{X}_{it}=\boldsymbol{x})}=\boldsymbol{x}'\boldsymbol{\gamma}_{\bar{h}h},\quad \bar{h},h=1,\ldots,k,\ h\neq\bar{h}. \tag{3}$$

Estimation of these LM models typically relies on the maximum likelihood method. Some details about this aspect are provided in the following, after having introduced the assumptions of the proposed CLT model.

Concluding this preliminary section, it is worth recalling that the use of discrete latent variables that characterizes LM models has certain advantages with respect to using continuous latent variables. Among these advantages, we can mention the flexibility, because with the proper number of latent states it is possible to approximate any continuous distribution adequately. Moreover, this approach is particularly useful when the interest is in clustering units in homogenous groups; within the LM approach this clustering is

dynamic, in the sense that the same unit can be assigned to different groups across time. For a deeper discussion on these points, see Bartolucci et al. (2022).

2.2 Model assumptions

In the following, we formulate the CLT model with explicit reference to two time occasions ($T = 1$), corresponding to the specific context of application of interest. Moreover, as in the standard DiD method, we make use of baseline covariates that are time-constant and are collected in the vectors $\boldsymbol{X}_i$. We assume that the individual-specific response variables depend on a vector $\boldsymbol{H}_i = (H_{i0}, H_{i1})'$ of two latent variables having a discrete distribution with support $\{1, \ldots, k\}$. Moreover, we assume conditional independence between the response variables given the latent process at different time occasions.

As mentioned above, we define a specific conditional distribution of the responses for each latent state. In our application, in particular, we rely on assumption (1), where the conditional means $\boldsymbol{\mu}_h$, $h = 1, \ldots, k$, and the variance-covariance matrix $\boldsymbol{\Sigma}$ are parameters whose estimates permit to interpret the latent states, as will be clear in Section 6. Obviously, the Gaussian distribution is a natural choice given that the test scores considered in the application are measured on a continuous scale. However, the present approach may be extended to deal with response variables having a different nature, even categorical, and then other distributions may be easily included; see Bartolucci et al. (2013).

We conceive the CLT model considering potential versions of the latent variables H_{it}. In particular, underlying every H_{it} we assume the existence of the potential latent variable $H_{it}^{(g)}$ corresponding to the latent state of individual i at occasion t if he/she had taken the treatment ($g = 1$) or not ($g = 0$). On the basis of these latent variables, we formulate the average treatment effect on the treated (ATET) measured on the logit scale. More importantly, this causal effect is specific of the two potential latent states at the two time occasions, that is,

$$\begin{aligned} \text{ATET}_{1\bar{h}h}(\boldsymbol{x}) = & \log \frac{p(H_{i1}^{(1)} = h | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)}{p(H_{i1}^{(1)} = \bar{h} | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)} \\ & - \log \frac{p(H_{i1}^{(0)} = h | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)}{p(H_{i1}^{(0)} = \bar{h} | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)}, \end{aligned} \tag{4}$$

where $\bar{h}$ is referred to the latent state at the first occasion and h at the second. Note that the above definition is conditional on a given value of the baseline covariates, denoted by $\boldsymbol{x}$, and it is referred to specific subpopulations. However, as will be clear in the following, when we formulate a suitable regression model for the latent POs we assume that the causal effect is constant with respect to $\boldsymbol{x}$.

We formulate the following assumptions to identify the above causal effects:

1. Stable Unit Treatment Value Assumption (SUTVA), according to which:

$$H_{it} = g_i H_{it}^{(1)} + (1 - g_i) H_{it}^{(0)}, \quad t = 0, 1;$$

   therefore, the outcome experienced by individual i is not affected by the assignment and received treatment by other individuals or, in other terms, there are no relevant interactions between members of the population.

2. Exogeneity (EXOGEN), according to which the covariates in $\boldsymbol{X}_i$ are time invariant and measured at the initial period before the treatment assignment or time variant but they are not influenced by the treatment.

3. No Effect for the Pretreatment Population (NEPT), which is motivated by the fact that the treatment is administrated between the two occasions and, therefore, it has no effect at $t = 0$. Consequently, it results that

$$p(H_{i0}^{(1)} = H_{i0}^{(0)} | \boldsymbol{X}_i = \boldsymbol{x}, G_i = g) = 1, \quad g = 0, 1, \; \forall \, \boldsymbol{x} \in \mathcal{X}.$$

4. Common Support (COSU), according to which every individual has a positive probability of receiving any type of the treatment; it is also named positivity assumption.

5. Common Trend (CT), according to which, in terms of transition probabilities, we

have

$$\log \frac{p(H_{i1}^{(0)} = h | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)}{p(H_{i1}^{(0)} = \bar{h} | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)} = \log \frac{p(H_{i1}^{(0)} = h | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 0)}{p(H_{i1}^{(0)} = \bar{h} | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 0)},$$

for $\bar{h}, h = 1, \ldots, k,\ h \neq \bar{h},\ \forall\, \boldsymbol{x} \in \mathcal{X}$.

According to Lechner (2011), p. 179, and with reference to the DiD, the CT assumption states that: "the differences in the expected potential non-treatment outcomes over time (conditional on X) are unrelated to belonging to the treated or control group in the post-treatment period. This is the key assumption of the DiD approach. It implies that if the treated had not been subjected to the treatment, both subpopulations defined by $D = 1$ and $D = 0$ would have experienced the same time trends conditional on X." Note, in particular, that with references to the CLT model, this assumption is directly formulated on the transition probabilities from $H_{i0}^{(0)} = \bar{h}$ to $H_{i1}^{(0)} = h$, given the covariates, which may be interpreted on the same footing as differences between conditional expected values used to define CT in the standard DiD framework. Similarly to the differences between conditional expected values, the transition probabilities do not depend on the treatment group so that non-treated units represent a proper counterfactual. On the other hand, we allow the potential outcome for the initial occasion to depend on the group although, on the basis of NEPT, there is no difference between the two potential latent variables $H_{i0}^{(0)}$ and $H_{i0}^{(1)}$ because the treatment has not been administered yet. In this way, the proposed method also allows for a form of non-observable confounding as we do not require the potential outcomes to be conditionally independent of the treatment given the covariates as in other causal frameworks.

Apart from CT, an important condition of our approach is EXOGEN according to which the observed covariates, not related to the treatment, do not differently influence the treated and non-treated groups. Similar arguments hold for non-linear models where "the conditional expectation of the observable outcome variable is related to the conditional expectation of a latent outcome variable", by means of "a strictly monotonously increasing

and invertible function" (Lechner, 2011, pp. 200-203). As already mentioned, this is the case of the CLT model, where the link between the conditional expectations of observable outcomes and unobserved covariates is based in the logit function, which is one-to-one.

Now we can prove that the average effect of the treated group is identified. The NEPT assumption implies that Equation (4) can be rewritten as

$$\begin{aligned}\text{ATET}_{1\bar{h}h}(\boldsymbol{x}) = &\log \frac{p(H_{i1}^{(1)} = h | H_{i0}^{(1)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)}{p(H_{i1}^{(1)} = \bar{h} | H_{i0}^{(1)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)} \\ &- \log \frac{p(H_{i1}^{(0)} = h | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)}{p(H_{i1}^{(0)} = \bar{h} | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)}.\end{aligned}$$

Due to SUTVA, the first term of the previous equation is directly equal to

$$\log \frac{p(H_{i1} = h | H_{i0} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)}{p(H_{i1} = \bar{h} | H_{i0} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)},$$

whereas CT and SUTVA imply that the second term is equal to

$$\log \frac{p(H_{i1} = h | H_{i0} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 0)}{p(H_{i1} = \bar{h} | H_{i0} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 0)}.$$

In the end, it results that

$$\begin{aligned}\text{ATET}_{1\bar{h}h}(\boldsymbol{x}) = &\log \frac{p(H_{i1} = h | H_{i0} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)}{p(H_{i1} = \bar{h} | H_{i0} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)} \\ &- \log \frac{p(H_{i1} = h | H_{i0} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 0)}{p(H_{i1} = \bar{h} | H_{i0} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 0)}.\end{aligned}$$

To apply the approach in practice it is convenient to formulate a multinomial logit model of the following type for the initial probabilities for $g = 0, 1$:

$$\begin{aligned}&\log \frac{p(H_{i0}^{(0)} = h | \boldsymbol{X}_i = \boldsymbol{x}, G_i = g)}{p(H_{i0}^{(0)} = 1 | \boldsymbol{X}_i = \boldsymbol{x}, G_i = g)} = \\ &= \log \frac{p(H_{i0}^{(1)} = h | \boldsymbol{X}_i = \boldsymbol{x}, G_i = g)}{p(H_{i0}^{(1)} = 1 | \boldsymbol{X}_i = \boldsymbol{x}, G_i = g)} = \beta_{0h}^{(g)} + \boldsymbol{x}'\boldsymbol{\beta}_{1h}, \quad h = 2, \ldots, k, \end{aligned} \tag{5}$$

where $\beta_{0h}^{(g)}$ allows us to account for the difference between treated and non-treated groups

in the initial period; this assumption is in agreement with the NEPT. For the transition between states at the second time occasion, the following logistic model is assumed in agreement with CT for $g = 0, 1$:

$$\log \frac{p(H_{i1}^{(0)} = h | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = g)}{p(H_{i1}^{(0)} = \bar{h} | H_{i0}^{(0)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = g)} = \gamma_{0\bar{h}h}^{(0)} + \boldsymbol{x}' \boldsymbol{\gamma}_{1\bar{h}h}, \quad \bar{h}, h = 1, \ldots, k, \ h \neq \bar{h}. \tag{6}$$

We also assume that

$$\log \frac{p(H_{i1}^{(1)} = h | H_{i0}^{(1)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)}{p(H_{i1}^{(1)} = \bar{h} | H_{i0}^{(1)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 1)} = \gamma_{0\bar{h}h}^{(1)} + \boldsymbol{x}' \boldsymbol{\gamma}_{1\bar{h}h}, \quad \bar{h}, h = 1, \ldots, k, \ h \neq \bar{h}, \tag{7}$$

whereas the same logit referred to the probabilities $p(H_{i1}^{(1)} = h | H_{i0}^{(1)} = \bar{h}, \boldsymbol{X}_i = \boldsymbol{x}, G_i = 0)$ is left unspecified. Note that the covariates affecting the transition probabilities could also include the lagged response variables as we do in our application, illustrated in Section 6.

Parameters $\gamma_{0\bar{h}h}^{(0)}$ and $\gamma_{0\bar{h}h}^{(1)}$ may be interpreted in terms of causal effect of the treatment. In particular, for $h \neq \bar{h}$ we directly have that

$$\text{ATET}_{1\bar{h}h}(\boldsymbol{x}) = \delta_{\bar{h}h} = \gamma_{0\bar{h}h}^{(1)} - \gamma_{0\bar{h}h}^{(0)}; \tag{8}$$

as already mentioned, this effect is constant with respect to $\boldsymbol{x}$. We can easily express the causal effects on another scale. For instance, by taking the exponential of the expression in (8) we can express these effects as odds that are of more straightforward interpretation in certain fields. In addition, we can directly express these effects as differences between probabilities, as clarified in the following.

Finally, we make it clear that the parametrization on the initial and transition probabilities assumed in (5), (6), and (7) could be seen as restrictive. In particular, we could consider the case in which the regression coefficients for the covariates are group specific, not only the intercept. With reference to the transition probabilities, this amounts to include two separate vectors of coefficients, denoted by $\boldsymbol{\gamma}_{1\bar{h}h}^{(0)}$ and $\boldsymbol{\gamma}_{1\bar{h}h}^{(1)}$ for the non-treated and treated units, respectively. This implies a more complex way to define the $\text{ATET}_{1\bar{h}h}(\boldsymbol{x})$ and the overall causal effect with respect to that in (8). For this reason, we prefer to rely on the assumption that $\boldsymbol{\gamma}_{1\bar{h}h}^{(0)} = \boldsymbol{\gamma}_{1\bar{h}h}^{(1)} = \boldsymbol{\gamma}_{1\bar{h}h}$, with a similar restriction on

the initial probabilities. These restrictions can be checked in an application, as we will illustrate in Section 6; we also studied violations of these restrictions within the simulation experiments described in Section 4. Another possible extension to conceive is the presence of interactions between the covariates or the effect of suitable transformation of the covariates. In this case, however, it is sufficient to include such effects in the vectors $\boldsymbol{X}_i$ while retaining the same assumptions as above on the initial and transition probabilities.

2.3 Estimation

The previous assumptions, and in particular parametrizations (5), (6) and (7), imply the following reduced form for the initial and transition probabilities of the latent variables H_{it}:

$$\log \frac{p(H_{i0}=h|\boldsymbol{X}_i=\boldsymbol{x},G_i=g_i)}{p(H_{i0}=1|\boldsymbol{X}_i=\boldsymbol{x},G_i=g_i)} = \beta_{0h}^{(0)}+g_i\bar{\beta}_{0h}+\boldsymbol{x}'\boldsymbol{\beta}_{1h}, \quad h=2,\ldots,k, \tag{9}$$

$$\log \frac{p(H_{i1}=h|H_{i0}=\bar{h},\boldsymbol{X}_i=\boldsymbol{x},G_i=g_i)}{p(H_{i1}=\bar{h}|H_{i0}=\bar{h},\boldsymbol{X}_i=\boldsymbol{x},G_i=g_i)} = \gamma_{0\bar{h}h}^{(0)}+g_i\bar{\gamma}_{0\bar{h}h}+\boldsymbol{x}'\boldsymbol{\gamma}_{1\bar{h}h}, \quad \bar{h},h=1,\ldots,k,\ h\neq\bar{h}, \tag{10}$$

where $\bar{\beta}_{0h}=\beta_{0h}^{(1)}-\beta_{0h}^{(0)}$ and $\bar{\gamma}_{0\bar{h}h}=\gamma_{0\bar{h}h}^{(1)}-\gamma_{0\bar{h}h}^{(0)}=\delta_{\bar{h}h}$ corresponds to the $\text{ATET}_{1\bar{h}h}(\boldsymbol{x})$ according to (8). These two equations for the first time occasion and for the transition between the two time occasions correspond to Equations (2) and (3), respectively, in the standard LM model with covariates.

Estimation is carried out on the basis of the maximum likelihood approach, as shown in Bartolucci et al. (2014). The likelihood function of the model is maximized through the EM algorithm (Baum et al., 1970; Dempster et al., 1977), where the manifest distribution of the observed responses is computed through suitable recursions (see Bartolucci et al., 2013, Ch.5, for details about its implementation). The algorithm alternates two steps until convergence: at the E-step we compute the expected value of the so-called complete data log-likelihood given the observed data and the current value of the parameters; at the M-step we maximize the expected complete data log-likelihood with respect to the model parameters, so we update the vector of parameters. These two steps are iterated

until convergence is reached.

Standard errors for the parameter estimates are obtained by exact computation of the information matrix or through reliable numerical approximations of this matrix. In our application, and as is rather common, we select the number of latent states (k) through the Bayesian Information Criterion (BIC, Schwarz, 1978), which typically leads to a more parsimonious model with respect to other selection criteria (Bacci et al., 2014). A detailed simulation study proposed in Bartolucci et al. (2016) shows the validity of this criterion also for the potential outcome formulation of the LM model.

Finally, note that it is also important to predict the sequence of latent states for a given unit in the sample over time. In particular, *path prediction* corresponds to predicting the latent state for each time occasion given the observed data, and it is obtained on the basis of the posterior distribution of the latent variables. This procedure is also named *local decoding*.

Suitable procedures to properly initialize the EM algorithm and perform model selection, and other computational tools required for the estimation and prediction, are available in the `R` package `LMest` (Bartolucci et al., 2017).

3 Further details on the proposed approach

In this section we provide some comments about the proposed CLT approach, and we introduce some possible extensions.

3.1 Relevant features of the proposal

The CLT model addresses the following main issues:

(a) *Number and types of outcomes*: (i) the CLT model is formulated in a multivariate form, and it allows us to estimate different causal effects of the treatment by looking at the joint variability of the responses over time; (ii) the CLT model is a non-linear model that overcomes the problems related to the scale dependence and the limited support of the variables. The probability distribution of the potential latent variables given the treatment and the pre-treatment covariates is invariant with

respect to transformations of these variables. The observed responses are related to the latent variables by means of the distribution in Equation (1). Moreover, the CLT respects the identifiability conditions requested for a casual model by Puhani (2012).

(b) Instead of comparing multiple static models, the CLT approach allows studying the initial conditions through the estimates of the initial probabilities, and then it analyzes sequential stage developments through the estimated transition probabilities.

(c) *Unobserved heterogeneity*: in many cases, especially with big data, huge populations are composed of specific subpopulations that differ for unobserved characteristics. In such a situation, treatment may have a different effect on each subpopulation, and the DiD method cannot jointly measure all these effects. On the other hand, the CLT model allows us to detect unobserved heterogeneity differently with respect to the proposal of Keane and Wolpin (1997), and it also allows us to account for the potential endogeneity of latent abilities (Hansen et al., 2003) as well as to discover latent clusters on the basis of the observed outcomes. The number of these latent groups is not fixed a priori but it is suitably determined; see also the discussion in Section 7. The $\text{ATET}_{1\bar{h}h}(\boldsymbol{x})$ in (4) is measured for each pair of subgroups $(\bar{h}, h)$, with $\bar{h}, h = 1, \ldots, k$, and in this way it is possible to verify if the treatment has different impacts.

(d) In the CLT model the outcomes are only dependent on the latent POs, influenced by the observed pre- and post-treatment covariates. These latent variables are defined differently from the factorial model (Cunha et al., 2010) as they are assumed to follow a Markov process (Bartolucci et al., 2014). In cases of incomplete information, the proposal overcomes the identification problems highlighted by Jöreskog (1966) because "identification requires that the investigator specifies some features of the model" as well as the indeterminacy of scores (Vittadini, 1989).

3.2 Possible extensions

In formulating the CLT model, we adopt a convenient logit parametrization to express the ATET. We can also write these effects directly in terms of differences between probabilities as an alternative of (4). In particular, consider the effects

$$\begin{aligned}\text{ATET}^{*}_{1\bar{h}h}(\boldsymbol{x}) &= p(H_{i1}^{(1)}=h|H_{i0}^{(1)}=\bar{h},\boldsymbol{X}_i=\boldsymbol{x},G_i=1)\\ &\quad -p(H_{i1}^{(0)}=h|H_{i0}^{(0)}=\bar{h},\boldsymbol{X}_i=\boldsymbol{x},G_i=1),\end{aligned} \tag{11}$$

where, again, $\bar{h}$ is referred to the latent state at the first time occasion and h is that at the second occasion. Given assumptions (6) and (7), it is possible to express this effect as

$$\text{ATET}^{*}_{1\bar{h}h}(\boldsymbol{x})=\frac{\exp(\gamma^{(1)}_{0\bar{h}h}+\boldsymbol{x}'\boldsymbol{\gamma}_{1\bar{h}h})}{1+\sum_{h'\neq h}\exp(\gamma^{(1)}_{0\bar{h}h'}+\boldsymbol{x}'\boldsymbol{\gamma}_{1\bar{h}h'})}-\frac{\exp(\gamma^{(0)}_{0\bar{h}h}+\boldsymbol{x}'\boldsymbol{\gamma}_{1\bar{h}h})}{1+\sum_{h'\neq h}\exp(\gamma^{(0)}_{0\bar{h}h'}+\boldsymbol{x}'\boldsymbol{\gamma}_{1\bar{h}h'})}$$

for given $\boldsymbol{x}$ and $h\neq\bar{h}$, where the denominators are multinomial logit normalizing constants. The previous expression may be exploited to express an estimate of the ATET on the probability scale once the model parameters have been estimated.

It may be also of interest to express the causal effect of the treatment directly on the observable outcomes. In this case, for outcome of type j, $j=1,\ldots,r$, we have the effect expressed as

$$\begin{aligned}\text{ATET}^{\dagger}_{j}(\boldsymbol{x}) &= \sum_{h=1}^{k}\text{E}(Y_{ij1}|H_{i1}=h)p(H_{i1}^{(1)}=h|\boldsymbol{X}_i=\boldsymbol{x},G_i=1)\\ &\quad -\sum_{h=1}^{k}\text{E}(Y_{ij1}|H_{i1}=h)p(H_{i1}^{(0)}=h|\boldsymbol{X}_i=\boldsymbol{x},G_i=1),\end{aligned} \tag{12}$$

where Y_{ij1} is an element of $\boldsymbol{Y}_{i1}$ and

$$\begin{aligned}p(H_{i1}^{(g)}=h|\boldsymbol{X}_i=\boldsymbol{x},G_i=1) &= \sum_{\bar{h}=1}^{k}p(H_{i0}^{(g)}=\bar{h}|\boldsymbol{X}_i=\boldsymbol{x},G_i=1)\\ &\quad \times p(H_{i1}^{(g)}=h|H_{i0}^{(g)}=\bar{h},\boldsymbol{X}_i=\boldsymbol{x},G_i=1),\quad g=0,1,\end{aligned}$$

is the probability at the second time occasion that the potential latent outcome for treat-

ment g is equal to h. Even in this case the causal effects may be estimated on the basis of the parameter estimates by exploiting the previous formulae.

Finally, as already mentioned, the approach may be easily extended to deal with settings in which more than two occasions of observation are available. In this case the model will be based on an initial probability formulation of type (5) and a sequence of $T-1$ transition probabilities of type (6) and (7). Moreover, the treatment effects may be formulated for $t = 1, \ldots, T-1$ with expressions of type (4), (11), and (12), which are denoted by $\text{ATET}_{t\bar{h}h}(\boldsymbol{x})$, $\text{ATET}^{*}_{t\bar{h}h}(\boldsymbol{x})$, and $\text{ATET}^{\dagger}_{jt}(\boldsymbol{x})$, respectively.

4 Simulation study

In order to validate the proposed approach, we performed a simulation study related to the application presented in Section 5. This study is based on a benchmark design described in Section 4.1, whose results are commented in Section 4.2, and on alternatives to this design based on using a larger set of covariates and misspecified models presented in Section 4.3.

4.1 Benchmark design

For a sample of size n, with $n = 1,000$, $2,000$, we considered individual vectors of three exogenous covariates $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})'$, the first two of which are continuous and the third is dichotomous, and individual vectors of $r = 2$ response variables $\boldsymbol{Y}_{i0} = (Y_{i01}, Y_{i02})'$ and $\boldsymbol{Y}_{i1} = (Y_{i11}, Y_{i12})'$ for the two time occasions, with $i = 1, \ldots, n$. The covariates are generated by letting $X_{i1} = X^{*}_{i1}$, $X_{i2} = X^{*}_{i2}$, and $X_{i3} = 2 \cdot I(X^{*}_{i3} \geq 0) - 1$, where $I(\cdot)$ is the indicator function equal to 1 if its argument is true and 0 otherwise, with X^{*}_{i1}, X^{*}_{i2}, and X^{*}_{i3} having the following trivariate Gaussian distribution:

$$\begin{pmatrix} X^{*}_{i1} \\ X^{*}_{i2} \\ X^{*}_{i3} \end{pmatrix} \sim N_3 \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix} \right].$$

Data are generated from a model with $k = 2, 3$ latent states. Under this model, the response variables have a bivariate Gaussian distribution with mean depending on the latent state, denoted by $\boldsymbol{\mu}_h = (\mu_{h1}, \mu_{h2})'$ for latent state h, with values increasing with h. The conditional variance-covariate matrix $\boldsymbol{\Sigma}$ is common to all latent states and assumes values corresponding to different levels of correlation ρ. Values of these parameters are reported in Table 1 of the SI file.

The initial states referred to $H_{i0}^{(g)}$, for $g = 0, 1$, are drawn from the logistic model in Equation (5), whereas, for the transition to the state at the second time occasion, the logistic models in (6) and (7) are assumed, depending on parameters having specific values. Note that these values are chosen so that treated individuals tend to belong to the second (or third) latent state with higher probability at the beginning and that the treatment has a positive effect in terms of transition probabilities. Values of the parameters involved in these model components are again reported in Table 1 of the SI file.

Finally, the assignment of the treatment is based on the logistic model

$$\log \frac{p(G_i = 1|\boldsymbol{X}_i = \boldsymbol{x})}{p(G_i = 0|\boldsymbol{X}_i = \boldsymbol{x})} = \alpha_0 + \boldsymbol{x}'\boldsymbol{\alpha}_1,$$

with two possible values for the intercept, $\alpha_0 = -1, 0$, corresponding to two different proportions of treated and non-treated individuals, and two possible values for $\boldsymbol{\alpha}_1$, equal to $\mathbf{0}$ or $0.5 \cdot \mathbf{1}$, corresponding to the situation of exogenous or endogenous treatment, where $\mathbf{1}$ is a vector of ones of suitable dimension.

Overall, we considered 32 different scenarios, corresponding to the combination of two different values of n, ρ, k, α_0, and $\boldsymbol{\alpha}_1$. Under each scenario, we drew 1,000 samples from the assumed model, and for every sample, we estimated the parameters of the proposed CLT model with covariates, also obtaining the standard errors by using the asymptotic method. Estimates for two versions of the model are compared: in the first, the only covariate is the indicator variable for the treatment, which is a misspecified model given the data generation process. In the second, the covariates are also included further to this indicator variable; see the SI file for additional details.

4.2 Results under the benchmark design

In order to summarize the simulation results, we consider the average bias (Av.Bias) and the average root mean square error (Av.RMSE), which are computed as

$$\text{Av.Bias} \;=\; \frac{1}{kr}\sum_{h=1}^{k}\sum_{j=1}^{r}\left|\frac{1}{S}\sum_{s=1}^{S}(\hat{\mu}_{hj}^{[s]}-\mu_{hj}^{[0]})\right|,$$

and

$$\text{Av.RMSE} \;=\; \sqrt{\frac{1}{kr}\sum_{h=1}^{k}\sum_{j=1}^{r}\frac{1}{S}\sum_{s=1}^{S}(\hat{\mu}_{hj}^{[s]}-\mu_{hj}^{[0]})^2},$$

where $\hat{\mu}_{hj}^{[s]}$ denotes the estimate of μ_{hj} obtained for the s-simulated sample and $\mu_{hj}^{[0]}$ denotes its true value. Results in terms of these two indicators are reported in Table 2 of the SI file.

We conclude that the means of the latent states are properly estimated by both CLT models with and without covariates, apart from the indicator variable for the treatment. The bias and RMSE also behave as expected with respect to the sample size and the model complexity, with a typical decrease of both as the sample size (n) and the number of latent states (k) increase. In this regard, it is not possible to spot significant differences between the two methods.

Then we consider the estimation of the effects of main interest, which are the causal parameters $\delta_{\bar{h}h}$ defined in (8). When $k = 2$ with a binary treatment, the causal effects are two, whereas when $k = 3$ they are six. In this case, the simulation results are evaluated in terms of bias and root mean square error (RMSE) for every parameter. These results are reported in Tables 3 and 4 of the SI file.

We observe that the difference between the two methods is remarkable, with a clear advantage of the proposed approach that includes the covariates in the estimated model. This is particularly evident in terms of bias, which is low for the proposed model and severe when the model is estimated without covariates, even if the treatment is exogenous

given the covariates. The behavior of bias and RMSE with respect to n and k is as expected and coherent with the comments provided about Table 2 of the SI file.

Finally, we considered the precision of the method to obtain the standard errors for the parameter estimates in terms of relative bias (R.Bias), which is computed as

$$\text{R.Bias} = \frac{\frac{1}{S}\sum_{s=1}^{S}(\hat{se}(\hat{\delta}_{hj})^{[s]} - se(\hat{\delta}_{hj})^{[0]})}{se(\hat{\delta}_{hj})^{[0]}} - 1,$$

where $\hat{se}(\hat{\delta}_{hj})^{[s]}$ is the standard error for $\hat{\delta}_{hj}$ obtained on the basis of the s-th simulated sample and $se(\hat{\delta}_{hj})^{[0]}$ is its true value obtained as standard deviation of the $\hat{\delta}_{hj}^{[s]}$ parameter estimates. These results are reported in Table 5 of the SI file.

For the first 100 samples generated under the different scenarios of the benchmark design, we also performed the selection of the optimal number of states k on the basis of the BIC according to the same procedure that will be used in the application; see Section 6. We found that the correct number of states, equal to 2 or 3 depending on the specific scenario, has always been selected on the basis of this procedure.

4.3 Other simulations designs

As a first extension of the benchmark design illustrated above, we considered the case of a larger number of covariates, which is even closer to the context of the application. In particular, the simulation designs include seven additional covariates that are generated from independent Gaussian distributions with mean equal to the mean of the first three variables and variance equal to 1; in symbols, we have

$$X_{ij}|X_{i1}^* = x_{i1}^*, X_{i2}^* = x_{i2}^*, X_{i3}^* = x_{i3}^* \sim N\left[\frac{1}{3}(x_{i1}^* + x_{i2}^* + x_{i3}), 1\right], \quad j = 4, \ldots, 10.$$

The full vector of covariates, which now has dimension 10, is still denoted by $\boldsymbol{X}_i$ and is used both in the generation of the data and the treatment as described in Section 4.1. In particular, for all model components, the simulation models rely on the same values of the intercepts, while the vectors of regression coefficients are augmented with all elements equal to 0; see also Table 1 of the SI file. The full vector of these covariates is also used

in the model estimated for every simulated sample. Overall, the added covariates do not have a significant effect but, being highly correlated with the significant covariates, their presence may represent a challenge for the proposed approach.

The results of the additional simulation study described above show that, while the quality of the estimates of the μ_{hj} parameters is not affected by the higher number of covariates with respect to the benchmark design, the quality of the estimates of the initial and transition probabilities and, consequently of the $\delta_{\bar{h}h}$ parameters, worsens. With $n = 1,000$, in particular, for certain simulated samples the estimates of these regression parameters tend to extreme values and directly affect the Bias and RMSE. With $n = 2,000$ these extreme estimates are not observed, and the estimation results are overall rather similar to those obtained under the benchmark design. The main conclusion of this additional simulation scenario is that the approach must be carefully applied when there are many covariates and it is necessary to adopt an accurate selection of the covariates so as to avoid unreliable parameter estimates.

As outlined at the end of Section 2.2, the proposed approach assumes that the effect of the covariates on the transition probabilities is the same for non-treated and treated units, that is, $\boldsymbol{\gamma}^{(0)}_{1\bar{h}h} = \boldsymbol{\gamma}^{(1)}_{1\bar{h}h}$. We can then consider the implication of the violation of this assumption. For this aim we generated samples from a model that is similar to that used within the benchmark design with the main difference that the transition probabilities for non-treated units are computed as in (6) with a specific vector $\boldsymbol{\gamma}^{(0)}_{1\bar{h}h}$ and, similarly, those of the treated units are computed as in (7) with a specific vector $\boldsymbol{\gamma}^{(1)}_{1\bar{h}h}$. The assumed values of these new parameters within the simulation study are obtained as $\boldsymbol{\gamma}^{(0)}_{1\bar{h}h} = \boldsymbol{\gamma}_{1\bar{h}h} - 0.25 \cdot \mathbf{1}$ and $\boldsymbol{\gamma}^{(0)}_{1\bar{h}h} = \boldsymbol{\gamma}_{1\bar{h}h} + 0.25 \cdot \mathbf{1}$, with $\boldsymbol{\gamma}_{1\bar{h}h}$ having elements indicated in Table 1 of the SI file.

The results of this additional simulation scenario are very close to those obtained under the benchmark design in terms of Bias and RMSE of the estimators of the causal parameters of interest. In particular, the proposed CLT approach maintains a considerable advantage over the LM model without covariates in estimating these effects.

5 Application

A large literature demonstrates that there are strong links between NCSs and CSs both in the educational process and work environment (see, among others, Cunha et al., 2006; Heckman et al., 2006; Cunha and Heckman, 2007, 2008; Cunha et al., 2010; Heckman and Kautz, 2012; Heckman et al., 2014; OECD, 2015; West et al., 2016). Three studies are particularly relevant from the methodological point of view. Based on a static factor model, the first shows that CSs and NCSs are equally crucial to success in many life dimensions such as education, income level, employment, and adolescent "risky" behaviors (Heckman et al., 2006). The second study defines CSs and NCSs as unobservable traits generating observed outcomes such as learning test results, level of education, educational achievement, salary level, and performance in job career (Cunha et al., 2010). The mutual influence in causal terms of NCSs and CSs is assessed by accounting for the socio-economic characteristics of the family through a dynamic factor model. Edin et al. (2022) show that the economic return to the NCSs is higher than the return to CSs. Other researchers attempt to verify whether appropriate educational projects conceived to improve NCSs also improve CSs. See, among others, Tierney et al. (1995), Kahne and Bailey (1999), Martins (2010), Holmlund and Silva (2014), and García-Pérez and Hidalgo-Hidalgo (2017). In general, the current literature shows that the implemented tutoring and accompaniment activities decrease the dependence on drugs or alcohol. At the same time, the improvement of self-concept and school outcomes is minor, especially for those students with more critical family and social conditions.

Concerning our application, we address the following scientific question already introduced in Section 1: "Do NCSs programs causally determine an improvement of the CSs in the Italian educational context?". We use the proposed CLT model to evaluate whether programs that stimulate NCSs also lead to improvement in CSs, and we compare the effects with those estimated with the DiD method. First of all we describe the available data, particularly regarding the outcomes, the kind of NCSs considered, the other covariates, and the educational programs finalized to improve the NCSs.

In particular, the data concern a sample of primary and middle class students of the

PAT observed from the 5th grade through the 8th grade during the 2015-2018 school years. As previously indicated, 25 schools with 1,561 pupils (out of 77 with a total population of 5,502 students) freely accepted participating in the PAT survey in 2015. Among these schools, 12 (with 845 students in 111 classes) freely adopted the above mentioned educational programs to improve the NCSs. The data are derived by integrating five datasets illustrated at the beginning of Section 2 of the SI file. The PAT is an Italian region whose students show excellent test results and in which there are no severe socio-economic problems and attention to NCSs is already an established practice. In this way, the analysis of the link between NCSs and CSs is not affected by disturbing factors.

5.1 Outcomes

The measurement of the CSs that are the outcomes of our analysis is based on standardized national tests. In fact, we consider the scores students achieved in the INVALSI tests in the 5th and 8th grades (primary outcomes). The tests are explicitly built to assess the students' knowledge of Italian literacy and Mathematics nationwide and are carried out with different degrees of difficulty and methods. For example, in the 5th grade (elementary school), they are written on paper, and in the 8th grade (middle school), they use an adaptive computer technology. The observed score is obtained by counting the number of correct answers in the total: the achievement of 55-60% of correct answers on all tests certifies the sufficiency. The percentage of correct answers is reported net of cheating, to provide data as accurate as possible, a phenomenon detected through a statistical control referring to those "improper" behaviors held during the administration of the INVALSI tests (correct answers provided because copied from other students or books or even suggested more or less explicitly by teachers). The national average of the test scores on the Rasch scale for each grade is fixed at 200. Data sources and descriptive statistics on these variables are in Section 2 of the SI file through Tables from 6 to 10.

5.2 Description of the non-cognitive skills and other covariates

The NCSs considered in our analysis are five main distinct but related personality traits, named the Big Five (John et al., 1999; Heckman et al., 2014), which correspond to the following five dimensions: (*i*) openness to experience, namely the propensity to open oneself to reality and new cultural or intellectual experiences; (*ii*) conscientiousness, namely the disposition to be responsible, hardworking, and organized; (*iii*) extraversion, namely the openness of oneself toward other people and things at the origin of a general behavior in living class and school education activities; (*iv*) agreeableness, namely the orientation towards cooperation, altruism, and cordiality in social relations that generates personal and social level of human and friendly relationships between students and among students and teachers for what concerns the school environment; (*v*) emotional stability (or in opposite meaning neuroticism), namely the containment of the emotional reactions, without sudden mood changes. Some other NCSs are (*vi*) school motivation, namely students' desire to participate in learning activities to improve knowledge; (*vii*) external locus of control, that is, the help that students need to achieve school goals (Gagné and Deci, 2005). Table 6 of the SI file describes in more detail non-cognitive skills considered in the illustrative example. The covariates are selected according to substantive knowledge of the context and data and considering the recent literature on the topic, as illustrated in the previous section.

5.3 Educational programs

Starting from 2015 up to 2018, the PAT elaborated plans for schools focused on student learning and the NCSs improvements involving teachers, active teaching methodologies, information orientation, training, and counseling. Very solid activities were proposed, at several occasions, from a scientific and organizational point of view. They were structured and designed according to the following four macro-categories: (*i*) training orientation managed by teachers during school hours, inside the programs of disciplines-subjects of study, or inside the curriculum (for example, alternation of experiences school-work, etc.); (*ii*) counseling out of school hours generally managed by external experts to the

school; (*iii*) information and orientation including activities addressed to the whole school such as open days, orientation fairs, meeting with privileged witnesses, etc.; (*iv*) mixed projects (derived as combinations of the previous three activities), for example, projects to combat the risk of discomfort and early school leavers. These activities involved an information part (the school paths in the second cycle), a part of counseling (discovery and strengthening the identity of the students), and training activity of the teachers in reducing the risk of dropping out for the students. The schools themselves freely decided whether or not to carry out these training projects by communicating their choice to the PAT. Once a school chooses to participate, all students compulsorily participate in the same activities with the same time commitment that varies among projects. From the institutional point of view, the schools were allowed to: (*i*) implement their own projects concerning their actual educational offer, including special activities for students; these were carried out even with some involvement with local authorities, and frequently they were out of school; and (*ii*) choose improvement projects from a list of projects proposed by the PAT and related to the students' learning objectives (INVALSI, academic achievement, skills certifications, etc.).

5.4 Absence of self-selection and check of causal latent transition model assumptions

In dealing with an observational study, we have to exclude any self-selection both of the schools participating and non-participating to the survey and of the schools participating and non-participating to the educational PAT programs. First of all, we examined whether the schools that voluntarily participated in the survey had students who, on average and with reference to the INVALSI 2015 test, had the same level of cognitive ability as students in schools that did not participate. We consider the average achievement scores in Italian and Mathematics for each school, and we compared participating and non-participating schools according to such average scores. We recall that there are 25 participating schools with 1,561 pupils and 52 non-participating schools with 3,941 pupils, and we recall that the school freely decides to adhere to the programs. The results

reported in Table 9 of the SI file show no significant differences between these two school types, so we can conclude for an absent or limited impact of self-selection.

We consider the average test scores in Italian and Mathematics in the 5th and 8th grades and average values for the covariates at the baseline ($t = 0$, 5th grade) across treated and non-treated students of the participating schools. According to the t-tests reported in Table 10 of the SI file, no significant differences either in Italian nor in Mathematics can be detected between the average scores of treated and non-treated students at the baseline ($t = 0$). We observe that at the baseline, students who received the treatment show an average score that is worse in Mathematics with respect to the score of not-treated students, while they have a better average score in Mathematics after the treatment. Treated students show higher values for all covariates: school motivation, quality of class relations, external support for student autonomy, well-being at school, discomfort at school, bullying acted, and bullying right away. The parental socio-economic status related to the international socio-economic index named ESCS is equal in both groups, although the parents' employment status is slightly higher for treated students. The proportion of females and students with fathers having an Italian nationality is similar between treated and non-treated students.

We can state that the model assumptions hold for this application, as explained in the following. First of all, SUTVA holds because either all classes in the same school carried out the educational program to increase NCSs, and there are no interactions between students in treated and non-treated schools. Second, regarding EXOG we have to consider that the NCSs and the other covariates in the model are those collected before the treatment through the INVALSI 2015 test, such as social capital and socio-economic and demographic characteristics of the students; thus they are not influenced by the treatment. Third, NEPT holds because the PAT educational programs were implemented between the 5th and 8th grades without affecting the previous individual characteristics. In effect, in the present case, the observed covariates are defined before the beginning of the treatment, and therefore they are time invariant. Moreover, the results in Table 10 of SI file show that the average values of the treated and non-treated outcomes both at times $t = 0$ an $t = 1$ are almost equal. As we show in Section 6, we can verify the effect

of the treatment only on the worst and best subgroups of students but not on the overall groups of the treated and non-treated students. This a posterior evidence that the CT property is respected. Fifth, COSU holds since treatments are considered conditional to the covariates in every group of schools, and each student has a positive probability of receiving the treatment. Therefore, the ATET, that is, the causal effect of the educational programs aimed to increase NCSs on CSs, may be identified.

6 Empirical results

First we show the results obtained with the proposed multivariate CLT model, and then, as a comparison, we also show the results obtained with the DiD approach as mentioned in Section 1. We account for missing values on the covariates through dummies as indicators for missing values (Dardanoni et al., 2011) in both models.

6.1 Results of the causal latent transition model

The CLT model is estimated as mentioned in Section 2.3 through the EM algorithm. Table 1 shows the results of the model selection procedure. The BIC index leads to selecting a model with two latent states.

Table 1: *Maximum log-likelihood, number of parameters, and BIC index for an increasing number of latent states ranging from 1 to 4*

k	$\hat{\ell}$	#par	BIC
1	-30647.51	5	61331.79
2	-30084.10	60	60609.39
3	-29890.94	147	60862.78
4	-29756.70	266	61469.32

According to the estimated conditional means shown in Table 2, which are increasingly ordered, we identified two subpopulations of students clustered in low and high levels of performance. Students in the first state or cluster show an average score of around 195 for both Italian and Mathematics, whereas students in the second cluster are the best performing since they show an average score of around 235 for Italian and 246 for

Table 2: *Estimated cluster conditional averages*

	Latent state (h)	
Scores	1	2
Italian	192.312	235.133
Mathematics	196.071	246.636

Mathematics, with an average gain of around 40 points on both subjects. It is worth mentioning that it is always possible to interpret the states as different achievement stages in subjects even under a model with more than two latent states. In fact, states can always be ordered according to the estimated conditional means, thus providing a proper interpretation as achievement levels.

At the beginning of the 5th grade, the average probability of belonging to the first cluster is 0.648. According to the estimated variance-covariance matrix, which is assumed as homogeneous across clusters, there is a weak positive association ($\hat{\rho} = 0.381$) between Italian literacy and achievement on Mathematics. Figure 1 shows the contour plot of the estimated marginal distributions. As we explain above, the average Italian score for both standardized tests is 200; therefore, students of the PAT region classified in the first cluster slightly underperform with respect to the national average and those in the second cluster are very well-performing.

Table 4 shows the effects of the covariates (described in Table 7 of the SI file) on the initial probabilities as in Equation (9). In 5th grade, females tend to belong to the cluster grouping students with top performance levels: the odds ratio for females versus males is equal to $\exp(0.462) = 1.587$, thus showing higher CSs than males. Discomfort at school negatively affects cognitive performance, and the estimated log-odds ratio for distressed versus happy students is equal to 0.249, revealing the importance of this feeling. The parent's employment status and their Italian nationality appear to be important factors contributing to competitive advantages in terms of CSs for the students. Table 7 displays the estimated ATET and the effect of the covariates, among which we included the lagged response variables, thus relaxing the conditional independence assumption and excluding the covariates related to well-being collected in 2015. Regarding the logistic regression

Table 3: *Estimated variance-covariance matrix between Italian and Mathematics achievement scores*

Scores	Italian	Mathematics
Italian	840.30	319.53
Mathematics	319.53	841.90

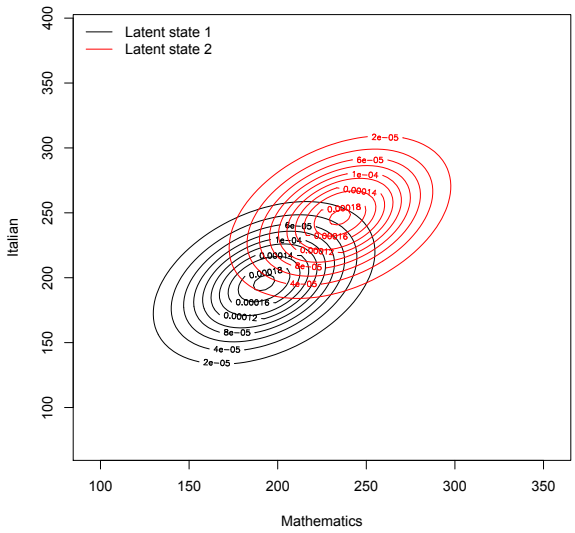


Figure 1: *Contour plot of the estimated densities for the two latent clusters according to the scores in Italian and Mathematics*

model for the transition probabilities, as in Equation (10), the average transition matrices are shown in Tables 5 and 6, whose standard errors are obtained through a non-parametric bootstrap based on 1,000 bootstrap samples. The first one shows a probability of around 0.342 of performing better in advanced studies, that is, of moving from the first to the second cluster. However, a similar probability (0.401) is estimated for moving from the second to the first cluster. Looking at Table 6 we observe that, while the probability to transit to the second cluster is roughly the same for treated and non-treated students, that from the second to the first cluster is higher for students who have not taken the educational programs aimed at improving the NCSs (0.463 versus 0.348). Therefore, non-treated students are more prone to worsening their CSs passing from the 5th to the 8th grade. The estimated ATET related to the transition from the second to the first cluster shown in Table 7 is negative and significant and the corresponding odds ratio for treated versus non-treated students is $\exp(-3.583) = 0.03$, showing that the proposed activities to improve NCSs reduces the probability that best students worsen during the school years. In 2015, females mainly belonged to the cluster of best performing students in both subjects but performed poorly over time compared to males (the coefficient related to the transition from first to the second cluster is negative and significant). Father's employment status is important, especially for moving from the first to the second group (the odds ratio is $\exp(0.394) = 1.483$). The log-odds of the Italian and Mathematics achievement scores at the fifth grade are positive for the transition from the first to the second cluster and negative for the transition from the second to the first cluster. Coherently with the value added theory (Bryk and Weisberg, 1976), what has been acquired in primary school helps to increase the CSs and reduces the possibility of decreasing cognitive abilities.

We performed some sensitivity analyses to validate the above results: (i) we investigated a possible differential treatment effect for Italian and Mathematics scores by estimating univariate models for each outcome; the results reported in Table 11 of the SI file confirm those obtained with the multivariate CTL model; (ii) we evaluated the plausibility of the conditional Gaussian distribution of each outcome once local decoding, mentioned at the end of Section 2.3, has been applied. In the SI file we show in Figures 1 and 2 the empirical conditional cumulative distribution functions for both outcomes in each

cluster at each grade; (*iii*) we checked also the results of the model with three clusters. We observe that these results are coherent with the previous results: the latent states are ordered for increasing values of the outcomes and they are in line with the results of the model with two latent states especially for what concerns the estimated treatment effects. We notice that according to the average transition matrices, non-treated students show a higher transition probability from latent state 2 to 1 and from latent states 3 to 2 compared to non-treated students. They also show a lower probability of remaining in latent states 2 and 3 compared to non-treated students; (*iv*) we estimated several models removing covariates in the initial and/or transition probabilities as well as considering some interaction effects of the treatment with other covariates such as gender, parents' socioeconomic index, and scores in Italian and Mathematics at grade 5th. These models showed a higher BIC index than that of the model reported above.

Finally, we have to stress that our analyses are valid under the CT assumption discussed in Section 2.2, which in general is a crucial assumption, in the DiD literature. Although we cannot perform a formal test on this assumption, we are confident it holds in the light of the data reported in Table 10 of the SI file. In fact, in this table we report the average score in Italian and Mathematics separately for treated and non-treated students, referred to 2012, when students were enrolled in the secondary elementary school year. Note that this period is earlier than the pretreatment year (2015). The comparison between the results for 2012 and 2015 leads to the conclusion that the CT is a realistic assumption.

6.2 Difference-in-Difference estimates

In the following, we report the results obtained with the standard DiD model expressed for the first time occasion as

$$Y_{i0} = \alpha + g_i\gamma + \boldsymbol{x}_i'\boldsymbol{\beta} + g_i\boldsymbol{x}_i'\boldsymbol{\phi} + \eta_{i0}, \tag{13}$$

Table 4: *Estimates of the logit regression parameters of the initial probability to belong to the second latent state with respect to the first latent state of the CLT model (significant *at 5%, **at 1%)*

Covariate	Effect	s.e.
Intercept	-3.275**	0.455
School motivations	-0.027	0.112
Parents' ESCS index	0.170	0.133
Quality of class relations	-0.180	0.184
External support for student autonomy	0.143	0.177
Well-being at school	0.198	0.199
Discomfort at school	-1.389**	0.163
Bullying acted	-0.685	0.495
Bullying right away	0.149	0.275
Female	0.462**	0.199
Italian nationality of the father	0.904**	0.308
Employment status of the father	0.186**	0.064
Employment status of the mother	0.232**	0.054
Missing indicator for parents' ESCS index	0.684**	0.199
Missing indicator for gender	2.771*	1.683
Missing indicator for father's nationality	-0.301	1.101
Missing indicator for father's employment	1.075**	0.344
Missing indicator for mother's employment	-1.172**	1.291

Table 5: *Average transition probabilities of the CLT model and, in parenthesis, estimated standard errors obtained through non-parametric bootstrap*

	Latent state (h)	
$\bar{h}$	1	2
1	0.658 (0.028)	0.342 (0.028)
2	0.401 (0.059)	0.599 (0.059)

Table 6: *Average transition probabilities of the CLT model for treated and non-treated students and, in parenthesis, estimated standard errors obtained through non-parametric bootstrap*

		Latent state (h)	
Treatment	$\bar{h}$	1	2
Treated	1	0.654 (0.040)	0.346 (0.040)
	2	0.348 (0.084)	0.652 (0.062)
Non-treated	1	0.662 (0.044)	0.338 (0.044)
	2	0.463 (0.073)	0.536 (0.073)

Table 7: *Estimates of the logit regression parameters of the transition probabilities under the CLT model: first column (Effect 1) from the first to the second cluster, second column (Effect 2) from the second to the first cluster (significant *at 5%, **at 1%)*

Covariates	Effect 1	s.e.	Effect 2	s.e.
Intercept	-43.393**	0.507	45.543**	0.802
Treatment	0.847	0.814	-3.583**	1.683
Parents' ESCS index	0.173	0.561	0.069†	1.116
Female	-2.702**	0.873	2.499**	1.565
Italian nationality of the father	0.344	0.833	4.425	2.881
Employment status of the father	0.394†	0.236	-0.026	0.426
Employment status of the mother	-0.314	0.239	-0.291	0.312
Missing indicator for ESCS index	0.102	0.938	-4.699**	1.792
Missing indicator for gender	-2.520	1.849	-3.134	3.433
Missing indicator for father's nationality	-0.399	2.511	19.818**	1.881
Missing indicator for parent's employment	1.998*	1.962	-6.046*	2.927
Missing indicator for mother's employment	-4.534*	1.962	-2.245	3.336
Italian score at the 5th grade	0.078**	0.011	-0.076**	0.022
Math score at the 5th grade	0.114**	0.010	-0.141**	0.020

where η_{i0} are error terms having zero mean and constant variance and considering as response the difference between the outcomes on the two time occasions:

$$Y_{i1} - Y_{i0} = \delta^{(0)} + \boldsymbol{x}_i'\boldsymbol{\lambda}^{(0)} + g_i\delta + \bar{\eta}_{it}. \tag{14}$$

This model implies that $\text{ATET}_1 = \delta$, so that it is independent of the covariates and can be simply estimated by the method of least squares on the basis of the observed data. Both models are estimated for the test results in Italian and Mathematics. Regarding the formulation in (13), we included all the available covariates, whereas in (14) we also added the previous achievement score, and we excluded covariates collected in 2015 related to the well-being at school.

Apart from the standard DiD formulation described above, we also considered the doubly robust estimator proposed by Sant'Anna and Zhao (2020) that, from a certain point of view, may be seen as a generalization of the DiD estimators proposed by Heckman et al. (1997) and Abadie (2005). In particular, we used the `R` package `DRDID` (Sant'Anna and Zhao, 2020) to estimate the models again for Italian and Mathematics scores separately.

In Tables 8 and 9 we show the estimated regression coefficients of the DiD models according to (13) (top panel) and (14) (bottom panel) without interactions between covariates. In order to better characterize some differences, we also provide the results of the models estimated with data of two subgroups of students: that of students with a test score above and below the median value at the 5th grade for both subjects.

Regarding the DiD models estimated assuming formulation (14), see the bottom panel of Table 8, females perform worse in Italian with respect to males and the family background is important to determine the student's performance: the estimated partial regression coefficients of the parents' ESCS index are positive and significant for Model 1 and Model 3 and for all the three models, respectively, see the caption of the tables for a description of each model. The programs to improve NCSs are effective only to improving score in Mathematics for Model 1 and Model 3. The coefficient related to the previous achievement is negative contrary to what is expected for Italian scores under Models 1 and 2, and for Mathematics scores under all the three models.

For the double robust DiD estimator proposed by Sant'Anna and Zhao (2020), the results are reported in Section 3.2 of the SI file (see Table 12). As for the other DiD models, also with this estimator, the treatment is not significant for Italian, while it is significant for Mathematics under Models 1 and 2.

Table 8: *Estimates of the regression parameters of the DiD models for Italian scores, as in Equation (13) (top panel) and Equation (14) (bottom panel), estimated for the overall students (Model 1), for the best performing students (Model 2), and for the worst performing students (Model 3) (†significant at 10%, *significant at 5%, **significant at 1%)*

Covariate	Model 1	Model 2	Model 3
Intercept	180.514**	225.749**	170.593**
Treatment	0.797	-0.419	0.635
School motivations	-0.301	1.235	-1.790†
Parents' ESCS index	-0.151	3.248*	-2.071*
Quality of class relations	-0.610	1.968	1.720
External support for student autonomy	2.618	3.557	0.402
Well-being at school	0.034	-1.689	0.463
Discomfort at school	-9.733**	-5.615**	-3.780**
Bullying acted	-7.855†	-5.596	-3.594
Bullying right away	1.400	2.859	0.119
Female	-4.081*	-0.165	-3.792*
Italian nationality of the father	13.225**	3.087	7.573**
Employment status of the father	2.235**	1.111†	0.132
Missing indicator for parents' ESCS index	5.291**	1.810	1.855
Missing indicator for gender	-4.407	7.715	-9.573
Missing indicator for father's nationality	9.313	-5.207	11.176†
Missing indicator for father's employment	13.247**	7.158*	4.098
Missing indicator for mother's employment	6.744	-2.261	8.475

Covariate	Model 1	Model 2	Model 3
Intercept	82.918**	83.649**	81.472**
Treatment	-0.426	-0.614	0.105
Parents' ESCS index	2.135*	1.208	4.973*
Female	-8.154**	-8.448**	-8.113**
Italian nationality of the father	0.235	1.245	-4.001
Employment status of the father	1.322**	0.914**	2.749**
Employment status of the mother	0.991*	1.186*	0.475
Missing indicator for parents' ESCS index	0.944	0.422	2.812
Missing indicator for gender	5.765	2.434	13.993
Missing indicator for father's nationality	-12.681*	-15.429†	-10.424
Missing indicator for father's employment	6.402**	5.061*	11.359*
Missing indicator for mother's employment	0.677	12.518	-20.520
Italian score at the 5th grade	-0.401**	-0.402†	-0.401†

Table 9: *Estimates of the regression parameters of the DiD models for Mathematics scores, as in Equation* (13) *(top panel) and Equation* (14) *(bottom panel), estimated for the overall students (Model 1), for the best performing students (Model 2), and the worst performing students (Model 3) (†significant at 10%, *significant at 5%, **significant at 1%)*

Covariate	Model 1	Model 2	Model 3
Intercept	190.214**	234.950**	176.686**
Treatment	-1.546	-2.145	-0.091
School motivations	-1.858	1.301	-2.400*
Parents' ESCS index	1.511	1.301	0.617
Quality of class relations	0.176	0.383	0.355
External support for student autonomy	-1.635	-4.166†	-0.434
Well-being at school	4.990*	0.801	3.602*
Discomfort at school	-12.362**	-4.625**	-4.147**
Bullying acted	-2.585	3.334	0.267
Bullying right away	-0.902	1.462	-3.040
Female	7.052**	6.789**	1.580
Italian nationality of the father	10.003**	0.735	5.087*
Employment status of the father	1.485*	1.410*	0.243
Employment status of the mother	1.742*	0.784	0.458
Missing indicator for parents' ESCS index	6.526**	2.311	0.761
Missing indicator for gender	28.446*	-38.052	2.659
Missing indicator for father's nationality	1.412	-6.096	9.082
Missing indicator for father's employment	6.014†	3.969	0.978
Missing indicator for mother's employment	-10.701	47.940†	2.048

Covariate	Model 1	Model 2	Model 3
Intercept	64.766**	67.157**	57.070*
Treatment	3.217*	2.542	5.437†
Parents' ESCS index	0.087	-0.636	1.960
Female	-0.789	-1.375	0.656
Italian nationality of the father	3.383	5.164*	-2.267
Employment status of the father	1.244*	1.094†	1.338
Employment status of the mother	1.049*	0.764	1.802*
Missing indicator for parents' ESCS index	2.619†	1.983	5.042
Missing indicator for gender	8.896	9.831	-4.698
Missing indicator for father's nationality	-12.275†	-8.880**	-19.789
Missing indicator for father's employment	8.671**	8.927	6.203
Missing indicator for mother's employment	-6.969	-9.306	8.449
Mathematics score at the 5th grade	-0.342**	-0.354**	-0.296**

7 Conclusions

We propose a Causal Latent Transition (CLT) model to estimate a treatment effect when observations are collected at two time occasions, before and after the treatment. The model may be cast in the class of latent (hidden) Markov models and may be seen as an alternative to the Difference-in-Difference method when multivariate outcomes are of interest and heterogeneous causal effects may be associated with different subpopulations not directly observable.

In more detail, the main issues of the proposed approach are the following:

(a) it allows us to detect unobserved heterogeneity, account for the potential endogeneity of latent abilities (Hansen et al., 2003), and discover latent clusters, whose number is not known a priori. The causal effect is measured for each pair of subgroups and, in this way, it is possible to verify if the treatment has different impacts;

(b) the CLT approach is formulated as a multivariate non-linear model allowing the estimation of different causal effects by looking at the joint variability of the responses over time; the probability distribution of the potential latent variables given the treatment and the pre-treatment covariates is invariant with respect to transformations of these variables;

(c) rather then comparing multiple static models, the CLT approach analyzes the initial conditions by estimating the initial probabilities and sequential stage developments by estimating the transition probabilities;

(d) in the CLT model, the outcomes are only dependent on latent POs, which are influenced by the observed pre- and post-treatment covariates. Differently from factor analysis, they are assumed to follow a Markov process and, in this way, the identification problems and indeterminacy of scores that typically arise in the factor model are avoided.

Note that within the proposed approach, the number of causal estimands varies with the selected number of latent states, while in more traditional causal approaches the number of estimands is fixed. In this regard, selecting the number of states is a crucial

point and, apart from criteria based on the observed data (see, for instance Figure 1 of the Supplementary Information), this selection can be driven by reasons of interpretability depending on the specific application.

The proposal is illustrated by an extensive simulation study and an application to assess the educational programs aimed to improve non-cognitive skills on pupils. This effect is evaluated by considering the pupil's cognitive skills measured through standardized national tests in Italian and Mathematics administered in the 5th and 8th grades. We infer a positive effect of the treatment on the subgroup of pupils having higher cognitive abilities. The results have been validated through suitable sensitivity analyses.

Apart from the present application, the proposal can be suitable to analyze data with multiple outcomes deriving from many other observational studies where it is important to verify differential results of the effects of the treatment on heterogeneous populations. We notice that even if the assumptions of the current CLT model are formulated for two time periods, these may be simply generalized to the case of more time occasions, and our proposal can be valuable with panel data as well. Another possible extension would be to account for more levels in the data structure, such as to capture the school or the class effects, and therefore a multilevel model would result. The CLT may be formulated similarly to the model proposed in Bartolucci et al. (2011), where an additional discrete latent variable is considered to capture the cluster effect. Further extensions can be conceived using a probit link function and assuming an underlying continuous latent variable. However, a probit parameterization instead of the proposed logit formulation would imply a slightly more complex estimation procedure.

8 Acknowledgements

We greatly acknowledge the Trento Autonomous Provincial Authority (Provincia Autonoma di Trento, PAT) and the Italian National Institute for the Evaluation of the Educational System (INVALSI) for providing access to surveys and administrative data. G. Vittadini also thanks PAT for promoting and funding the present research through the ongoing work of IPRASE, the local institution devoted to research on education and

schooling.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72:1–19.

Bacci, S., Pandolfi, S., and Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, 8:125–145.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC, Boca Raton, FL.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2014). Latent Markov models: A review of a general framework for the analysis of longitudinal data with covariates (with discussion). *TEST*, 23:433–465.

Bartolucci, F., Pandolfi, S., and Pennoni, F. (2017). `LMest`: An `R` package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software*, 81:1–38.

Bartolucci, F., Pandolfi, S., and Pennoni, F. (2022). Discrete latent variable models. *Annual Review of Statistics and its Application*, 9:425–452.

Bartolucci, F., Pennoni, F., and Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, 36:491–522.

Bartolucci, F., Pennoni, F., and Vittadini, G. (2016). Causal latent Markov model for the comparison of multiple treatments in observational longitudinal studies. *Journal of Educational and Behavioral Statistics*, 41:146–179.

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171.

Becker, G. S. (1994). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*. The University of Chicago Press (3rd edition), New York.

Bouveyron, C., Celeux, G., Murphy, T., and Raftery, A. (2002). *Model-Based Clustering and Classification for Data Science, with Applications in R*. Cambridge University Press, Cambridge.

Bryk, A. S. and Weisberg, H. I. (1976). Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, 1:127–155.

Cunha, F. and Heckman, J. (2007). The technology of skill formation. *American Economic Review*, 97:31–47.

Cunha, F. and Heckman, J. J. (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources*, 43:738–782.

Cunha, F., Heckman, J. J., Lochner, L., and Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, 1:697–812.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78:883–931.

Dardanoni, V., Modica, S., and Peracchi, F. (2011). Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics*, 162:362–368.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Edin, P.-A., Fredriksson, P., Nybom, M., and Öckert, B. (2022). The rising return to noncognitive skill. *American Economic Journal: Applied Economics*, 14:78–100.

Früuhwirth-Schnatter, S., Celeux, G., and C.P., R. (2019). *Handbook of Mixture Analysis*. CRC press, Boca Raton, FL.

Gagné, M. and Deci, E. L. (2005). Self-determination theory and work motivation. *Journal of Organizational Behavior*, 26:331–362.

García-Pérez, J. I. and Hidalgo-Hidalgo, M. (2017). No student left behind? Evidence from the programme for school guidance in Spain. *Economics of Education Review*, 60:97–111.

Hansen, K., Heckman, J. J., and Mullen, K. J. (2003). The effect of schooling and ability on achievement test scores. *ERIC: NBER Working Paper Series, No. 9881*, 1:1–73.

Heckman, J. J., Humphries, J. E., and Kautz, T. (2014). *The Myth of Achievement Tests: The GED and the Role of Character in American Life*. University of Chicago Press.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64:605–654.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19:451–464.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24:411–482.

Holmlund, H. and Silva, O. (2014). Targeting noncognitive skills to improve cognitive outcomes: Evidence from a remedial education intervention. *Journal of Human Capital*, 8:126–160.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47:5–86.

John, O. P., Srivastava, S., et al. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, 2:102–138.

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika*, 31:165–178.

Kahne, J. and Bailey, K. (1999). The role of social capital in youth development: The case of "I have a dream" programs. *Educational Evaluation and Policy Analysis*, 21:321–343.

Keane, M. P. and Wolpin, K. I. (1997). The career decisions of young men. *Journal of Political Economy*, 3:473–522.

Lanza, S. T., Coffman, D. L., and Xu, S. (2013). Causal inference in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 20:361–383.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, 4:165–224.

Lee, M.-J. (2016). *Matching, Regression Discontinuity, Difference in Differences, and Beyond*. Oxford University Press.

MacDonald, I. L. and Zucchini, W. (2016). Hidden markov models for discrete-valued time series. In RA Davis, SH Holan, R. L. and Ravishanker, N., editors, *Handbook of Discrete-Valued Time Series*, pages 267–286. Chapman and Hall/CRC, Boca Raton, FL.

Martins, P. S. (2010). Can targeted, non-cognitive skills programs improve achievement? Evidence from EPIS. Technical report, IZA Discussion Paper.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

OECD (2015). *Skills for Social Progress: The Power of Social and Emotional skills*. OECD Publishing.

Puhani, P. A. (2012). The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. *Economics Letters*, 115:85–87.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, 7:143–176.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Tierney, J. P. et al. (1995). Making a difference. An impact study of big brothers/big sisters. Technical report, ERIC.

Vittadini, G. (1989). Indeterminacy problems in the LISREL model. *Multivariate Behavioral Research*, 24:397–414.

Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, 53:1–13.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F., and Gabrieli, J. D. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis*, 38:148–170.

Supplementary information for the paper titled: "A Causal Latent Transition Model With Multivariate Outcomes and Unobserved Heterogeneity: Application to Human Capital Development"

Francesco Bartolucci
University of Perugia (IT)
francesco.bartolucci@unipg.it

Fulvia Pennoni
University of Milano-Bicocca (IT)
fulvia.pennoni@unimib.it

Giorgio Vittadini
University of Milano-Bicocca (IT)
giorgio.vittadini@unimib.it

We first provide some details on the simulation design presented in Section 4 of the paper, and we show additional simulation results to evaluate the proposed causal latent transition (CLT) model. In Section 2 we describe the data introduced in Section 5 of the paper referred to the empirical application and show summary statistics of the chosen variables. In Section 3 we show some additional results of the application proposed in Section 6 of the paper. An example code developed for the simulations and the application is available at the GitHub repository: `https://github.com/penful/CausalLT`.

1 Simulation design and results

In the following, we first report the details on the parameter values under the two simulated scenarios presented in Section 4.1 of the paper. Then, we provide tables with the complete results of the simulation study described in Section 4.2 of the paper.

1.1 Simulation design

Table 1 reports the values of the parameters of the benchmark design presented in Section 4.1 of the paper.

Table 1: *Parameter values under the different simulation scenarios for a causal latent transition model with 2 and 3 latent states*

$k = 2$	$k = 3$
$\boldsymbol{\mu}_1 = \begin{pmatrix} -1 \\ -2 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$	$\boldsymbol{\mu}_1 = \begin{pmatrix} -1.5 \\ -3 \end{pmatrix}$, $\boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\boldsymbol{\mu}_3 = \begin{pmatrix} 1.5 \\ 3 \end{pmatrix}$
$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, $\rho = 0.25, 0.75$	
$\beta_{02}^{(0)} = 0$	$\beta_{02}^{(0)} = 1$, $\beta_{03}^{(0)} = 0.25$
$\beta_{02}^{(1)} = 1$	$\beta_{02}^{(1)} = 2$, $\beta_{03}^{(1)} = 1.25$
$\boldsymbol{\beta}_{12} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$	$\boldsymbol{\beta}_{12} = \boldsymbol{\beta}_{13} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$
$\gamma_{012}^{(0)} = \gamma_{021}^{(0)} = 0$	$\gamma_{012}^{(0)} = \gamma_{013}^{(0)} = 0$, $\gamma_{021}^{(0)} = \gamma_{023}^{(0)} = -1$, $\gamma_{031}^{(0)} = \gamma_{032}^{(0)} = 0$
$\gamma_{012}^{(1)} = 1$, $\gamma_{021}^{(1)} = -1$	$\gamma_{012}^{(1)} = \gamma_{013}^{(1)} = 1$, $\gamma_{021}^{(1)} = -1$, $\gamma_{023}^{(1)} = 1$, $\gamma_{031}^{(1)} = \gamma_{032}^{(1)} = -1$
$\boldsymbol{\gamma}_{112} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$	$\boldsymbol{\gamma}_{112} = \boldsymbol{\gamma}_{113} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$, $\boldsymbol{\gamma}_{121} = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$
$\boldsymbol{\gamma}_{121} = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$	$\boldsymbol{\gamma}_{123} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$, $\boldsymbol{\gamma}_{131} = \boldsymbol{\gamma}_{132} = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$

1.2 Simulation results

Tables from 2 to 5 report the results of the simulations described in Section 4.2 of the paper.

Table 2: *Simulation results in terms of average bias (Av.Bias) and average root mean square error (Av.RMSE) for the μ_{hj} parameters. CLT refers to the CLT model in which the only covariate is the treatment, whereas CLTcov refers to the case in which the covariates are also included in the initial and transition logistic models*

							μ_{hj} $(k=2)$		μ_{hj} $(k=3)$	
n	ρ	α_0	α_{11}	α_{12}	α_{13}		Av.Bias	Av.RMSE	Av.Bias	Av.RMSE
1000	0.25	-1.00	0.00	0.00	0.00	CLT	0.0015	0.0339	0.0007	0.0505
						CLTcov	0.0014	0.0336	0.0007	0.0493
1000	0.25	-1.00	0.50	0.50	0.50	CLT	0.0011	0.0338	0.0011	0.0501
						CLTcov	0.0011	0.0335	0.0011	0.0490
1000	0.25	0.00	0.00	0.00	0.00	CLT	0.0007	0.0343	0.0018	0.0520
						CLTcov	0.0007	0.0341	0.0017	0.0502
1000	0.25	0.00	0.50	0.50	0.50	CLT	0.0013	0.0341	0.0010	0.0509
						CLTcov	0.0013	0.0338	0.0010	0.0496
1000	0.75	-1.00	0.00	0.00	0.00	CLT	0.0007	0.0335	0.0007	0.0497
						CLTcov	0.0005	0.0332	0.0006	0.0485
1000	0.75	-1.00	0.50	0.50	0.50	CLT	0.0010	0.0337	0.0020	0.0492
						CLTcov	0.0009	0.0334	0.0017	0.0484
1000	0.75	0.00	0.00	0.00	0.00	CLT	0.0005	0.0336	0.0008	0.0495
						CLTcov	0.0005	0.0333	0.0009	0.0484
1000	0.75	0.00	0.50	0.50	0.50	CLT	0.0005	0.0338	0.0010	0.0492
						CLTcov	0.0005	0.0335	0.0013	0.0479
2000	0.25	-1.00	0.00	0.00	0.00	CLT	0.0005	0.0239	0.0009	0.0357
						CLTcov	0.0005	0.0237	0.0009	0.0345
2000	0.25	-1.00	0.50	0.50	0.50	CLT	0.0006	0.0239	0.0008	0.0361
						CLTcov	0.0007	0.0237	0.0007	0.0351
2000	0.25	0.00	0.00	0.00	0.00	CLT	0.0006	0.0243	0.0012	0.0354
						CLTcov	0.0007	0.0241	0.0010	0.0345
2000	0.25	0.00	0.50	0.50	0.50	CLT	0.0005	0.0238	0.0010	0.0357
						CLTcov	0.0005	0.0237	0.0011	0.0346
2000	0.75	-1.00	0.00	0.00	0.00	CLT	0.0008	0.0239	0.0005	0.0347
						CLTcov	0.0008	0.0239	0.0004	0.0335
2000	0.75	-1.00	0.50	0.50	0.50	CLT	0.0007	0.0240	0.0011	0.0339
						CLTcov	0.0007	0.0237	0.0008	0.0327
2000	0.75	0.00	0.00	0.00	0.00	CLT	0.0005	0.0237	0.0007	0.0352
						CLTcov	0.0005	0.0235	0.0009	0.0342
2000	0.75	0.00	0.50	0.50	0.50	CLT	0.0005	0.0241	0.0008	0.0342
						CLTcov	0.0004	0.0240	0.0009	0.0332

Table 3: *Simulation results in terms of average bias for the $\delta_{\overline{hh}}$ parameters. CLT refers to the CLT model in which the only covariate is the treatment, whereas CLTcov refers to the case in which the covariates are also included in the initial and transition logistic models*

							Bias ($k=2$)		Bias ($k=3$)					
n	ρ	α_0	α_{11}	α_{12}	α_{13}		δ_{12}	δ_{21}	δ_{12}	δ_{13}	δ_{21}	δ_{23}	δ_{31}	δ_{32}
1000	0.25	-1.00	0.00	0.00	0.00	CLT	-0.2622	0.2618	-0.1435	-0.1374	0.2905	-0.3403	0.2269	0.1819
						CLTcov	0.0241	-0.0151	0.1200	0.0790	-0.0601	0.0115	-0.0550	-0.0978
1000	0.25	-1.00	0.50	0.50	0.50	CLT	-0.1820	0.2994	0.0107	-0.0028	0.2912	-0.3656	0.2420	0.2377
						CLTcov	0.0198	-0.0188	-0.0054	0.0309	-0.0376	0.0530	-0.0682	-0.0324
1000	0.25	0.00	0.00	0.00	0.00	CLT	-0.2568	0.2617	-0.1671	-0.1730	0.2998	-0.3423	0.2435	0.2383
						CLTcov	0.0330	-0.0122	0.0496	0.0500	-0.0532	0.0185	-0.0324	-0.0764
1000	0.25	0.00	0.50	0.50	0.50	CLT	-0.2603	0.3100	-0.0729	-0.1002	0.3599	-0.3495	0.2587	0.2674
						CLTcov	0.0137	-0.0010	0.0709	0.0334	-0.0361	0.0421	-0.0398	-0.0886
1000	0.75	-1.00	0.00	0.00	0.00	CLT	-0.2620	0.2542	-0.1506	-0.1167	0.2949	-0.3354	0.2441	0.2350
						CLTcov	0.0239	-0.0305	0.0618	0.0918	-0.0550	0.0173	-0.0494	-0.0692
1000	0.75	-1.00	0.50	0.50	0.50	CLT	-0.1894	0.2906	-0.0164	0.0050	0.2888	-0.3594	0.2638	0.2285
						CLTcov	0.0158	-0.0234	0.0309	0.0386	-0.0329	0.0450	-0.0513	-0.0970
1000	0.75	0.00	0.00	0.00	0.00	CLT	-0.2646	0.2701	-0.1675	-0.1585	0.3184	-0.3450	0.2457	0.2448
						CLTcov	0.0200	-0.0082	0.0345	0.0600	-0.0335	0.0031	-0.0465	-0.0541
1000	0.75	0.00	0.50	0.50	0.50	CLT	-0.2764	0.3082	-0.0902	-0.1051	0.3561	-0.3520	0.2584	0.2557
						CLTcov	0.0017	-0.0127	0.0655	0.0452	-0.0191	0.0292	-0.0562	-0.0612
2000	0.25	-1.00	0.00	0.00	0.00	CLT	-0.2650	0.2603	-0.1605	-0.1680	0.3238	-0.3356	0.2613	0.2040
						CLTcov	0.0141	-0.0119	0.0355	0.0234	-0.0192	0.0112	-0.0079	-0.0614
2000	0.25	-1.00	0.50	0.50	0.50	CLT	-0.1892	0.3009	-0.0264	-0.0173	0.2902	-0.3756	0.2777	0.2727
						CLTcov	0.0072	-0.0115	-0.0028	0.0109	-0.0270	0.0163	-0.0064	-0.0153
2000	0.25	0.00	0.00	0.00	0.00	CLT	-0.2711	0.2763	-0.1603	-0.1679	0.3349	-0.3338	0.2482	0.2426
						CLTcov	0.0074	0.0009	0.0450	0.0422	-0.0051	0.0123	-0.0190	-0.0247
2000	0.25	0.00	0.50	0.50	0.50	CLT	-0.2626	0.3029	-0.0890	-0.0982	0.3576	-0.3572	0.2829	0.2882
						CLTcov	0.0121	-0.0084	0.0283	0.0322	-0.0123	0.0157	-0.0153	-0.0098
2000	0.75	-1.00	0.00	0.00	0.00	CLT	-0.2779	0.2720	-0.1673	-0.1684	0.3138	-0.3460	0.2438	0.2297
						CLTcov	0.0048	0.0014	0.0343	0.0277	-0.0302	-0.0023	-0.0275	-0.0370
2000	0.75	-1.00	0.50	0.50	0.50	CLT	-0.1889	0.3023	-0.0243	-0.0306	0.2891	-0.3768	0.2850	0.2726
						CLTcov	0.0138	-0.0087	0.0025	0.0018	-0.0303	0.0127	-0.0119	-0.0187
2000	0.75	0.00	0.00	0.00	0.00	CLT	-0.2775	0.2700	-0.1655	-0.1828	0.3225	-0.3268	0.2484	0.2539
						CLTcov	-0.0016	-0.0028	0.0483	0.0332	-0.0213	0.0214	-0.0198	-0.0143
2000	0.75	0.00	0.50	0.50	0.50	CLT	-0.2652	0.2986	-0.0845	-0.1083	0.3474	-0.3671	0.2837	0.2549
						CLTcov	0.0118	-0.0100	0.0471	0.0198	-0.0277	0.0078	-0.0199	-0.0465

Table 4: *Simulation results in terms of root mean square error (RMSE) for the $\delta_{\bar{h}h}$ parameters. CLT refers to the CLT model in which the only covariate is the treatment, whereas CLTcov refers to the case in which the covariates are also included in the initial and transition logistic models*

							RMSE ($k=2$)		RMSE ($k=3$)					
n	ρ	α_0	α_{11}	α_{12}	α_{13}		δ_{12}	δ_{21}	δ_{12}	δ_{13}	δ_{21}	δ_{23}	δ_{31}	δ_{32}
1000	0.25	-1.00	0.00	0.00	0.00	CLT	0.3742	0.3324	0.8417	0.4952	0.4952	0.4254	0.4631	0.4996
						CLTcov	0.3054	0.2366	1.6893	0.4273	0.4273	0.2934	0.4619	0.5934
1000	0.25	-1.00	0.50	0.50	0.50	CLT	0.3112	0.3608	0.6033	0.4818	0.4818	0.4429	0.4632	0.6473
						CLTcov	0.2989	0.2676	1.1715	0.4259	0.4259	0.3471	0.5217	0.7017
1000	0.25	0.00	0.00	0.00	0.00	CLT	0.3379	0.3237	0.5413	0.4466	0.4466	0.4148	0.4295	0.4704
						CLTcov	0.2588	0.2152	0.6051	0.3699	0.3699	0.2667	0.3992	0.9735
1000	0.25	0.00	0.50	0.50	0.50	CLT	0.3405	0.3625	0.5295	0.4887	0.4887	0.4226	0.4465	0.4815
						CLTcov	0.2823	0.2493	0.6443	0.4161	0.4161	0.3065	0.4817	1.3158
1000	0.75	-1.00	0.00	0.00	0.00	CLT	0.3631	0.3282	1.0636	0.5119	0.5119	0.4205	0.4626	0.5124
						CLTcov	0.2962	0.2427	1.1400	0.4491	0.4491	0.2836	0.4496	0.7999
1000	0.75	-1.00	0.50	0.50	0.50	CLT	0.3151	0.3452	0.6086	0.4762	0.4762	0.4298	0.4487	0.4812
						CLTcov	0.3073	0.2534	0.7875	0.4438	0.4438	0.3323	0.4733	0.5507
1000	0.75	0.00	0.00	0.00	0.00	CLT	0.3445	0.3260	0.5228	0.4595	0.4595	0.4143	0.4223	0.4726
						CLTcov	0.2558	0.2145	0.8367	0.3679	0.3679	0.2657	0.3867	0.4583
1000	0.75	0.00	0.50	0.50	0.50	CLT	0.3522	0.3592	0.5109	0.4800	0.4800	0.4211	0.4280	0.4611
						CLTcov	0.2851	0.2531	0.6228	0.4055	0.4055	0.3081	0.4774	0.5234
2000	0.25	-1.00	0.00	0.00	0.00	CLT	0.3214	0.2980	0.4368	0.4202	0.4202	0.3802	0.3734	0.3917
						CLTcov	0.2097	0.1671	0.4645	0.2796	0.2796	0.2030	0.3029	0.3665
2000	0.25	-1.00	0.50	0.50	0.50	CLT	0.2626	0.3284	0.3936	0.3912	0.3912	0.4133	0.3853	0.4033
						CLTcov	0.2198	0.1838	0.4647	0.2962	0.2962	0.2339	0.3366	0.3815
2000	0.25	0.00	0.00	0.00	0.00	CLT	0.3162	0.3077	0.3928	0.4017	0.4017	0.3734	0.3506	0.3642
						CLTcov	0.1852	0.1521	0.4004	0.2460	0.2460	0.1888	0.2798	0.3132
2000	0.25	0.00	0.50	0.50	0.50	CLT	0.3055	0.3297	0.3647	0.4195	0.4195	0.3921	0.3708	0.3995
						CLTcov	0.2034	0.1766	0.4317	0.2772	0.2772	0.2195	0.3127	0.3704
2000	0.75	-1.00	0.00	0.00	0.00	CLT	0.3327	0.3076	0.4384	0.4067	0.4067	0.3835	0.3636	0.3695
						CLTcov	0.2047	0.1596	0.4687	0.2750	0.2750	0.1899	0.3116	0.3336
2000	0.75	-1.00	0.50	0.50	0.50	CLT	0.2620	0.3306	0.3991	0.3905	0.3905	0.4133	0.3845	0.3957
						CLTcov	0.2182	0.1768	0.4597	0.2884	0.2884	0.2275	0.3350	0.3665
2000	0.75	0.00	0.00	0.00	0.00	CLT	0.3180	0.3001	0.3783	0.3965	0.3965	0.3668	0.3497	0.3820
						CLTcov	0.1792	0.1512	0.3805	0.2473	0.2473	0.1928	0.2752	0.3164
2000	0.75	0.00	0.50	0.50	0.50	CLT	0.3055	0.3247	0.3508	0.4099	0.4099	0.4024	0.3756	0.3719
						CLTcov	0.1978	0.1736	0.4242	0.2778	0.2778	0.2200	0.3128	0.3564

Table 5: *Simulation results in terms of relative bias (R.Bias) for standard errors obtained for the $\delta_{\overline{hh}}$ parameters. CLT refers to the model in which the only covariate is the treatment, whereas CLTcov refers to the case in which the covariates are also included in the initial and transition logistic models*

							R.Bias s.e. ($k=2$)		R.Bias s.e. ($k=3$)					
n	ρ	α_0	α_{11}	α_{12}	α_{13}		δ_{12}	δ_{21}	δ_{12}	δ_{13}	δ_{21}	δ_{23}	δ_{31}	δ_{32}
1000	0.25	-1.00	0.00	0.00	0.00	CLT	-0.0137	0.0075	-0.2545	-0.2894	-0.0088	-0.0074	-0.0288	-0.0035
						CLTcov	-0.0106	-0.0089	-0.5717	-0.4135	-0.0069	-0.0239	-0.0452	-0.1040
1000	0.25	-1.00	0.50	0.50	0.50	CLT	-0.0092	-0.0255	-0.0155	-0.0223	-0.0327	-0.0299	-0.0501	-0.2807
						CLTcov	0.0044	-0.0235	-0.4068	-0.0253	-0.0117	-0.0330	-0.0651	-0.1943
1000	0.25	0.00	0.00	0.00	0.00	CLT	0.0047	-0.0340	-0.0321	-0.0564	-0.0208	-0.0089	-0.0172	-0.0201
						CLTcov	-0.0071	-0.0157	-0.0591	-0.0428	-0.0213	-0.0036	-0.0206	-0.5389
1000	0.25	0.00	0.50	0.50	0.50	CLT	0.0035	-0.0229	-0.0382	-0.0498	-0.0317	-0.0255	-0.0426	-0.0135
						CLTcov	0.0092	0.0002	-0.0319	-0.0633	-0.0253	0.0052	-0.0442	-0.5728
1000	0.75	-1.00	0.00	0.00	0.00	CLT	0.0385	-0.0160	-0.4131	-0.3327	-0.0730	-0.0238	-0.0146	-0.0270
						CLTcov	0.0092	-0.0358	-0.3811	-0.3190	-0.0758	-0.0081	-0.0340	-0.3851
1000	0.75	-1.00	0.50	0.50	0.50	CLT	-0.0118	0.0444	-0.0354	-0.0170	-0.0445	0.0055	0.0139	-0.0039
						CLTcov	-0.0287	0.0240	-0.1373	-0.0323	-0.0734	-0.0210	0.0112	-0.0092
1000	0.75	0.00	0.00	0.00	0.00	CLT	-0.0063	0.0001	-0.0162	0.0019	-0.0531	-0.0105	-0.0011	-0.0416
						CLTcov	-0.0097	-0.0220	-0.3397	-0.0321	-0.0519	-0.0272	0.0021	-0.0390
1000	0.75	0.00	0.50	0.50	0.50	CLT	0.0010	-0.0140	-0.0300	-0.0050	-0.0352	-0.0194	0.0039	0.0060
						CLTcov	-0.0071	-0.0199	-0.0353	-0.0203	-0.0338	-0.0278	-0.0483	-0.0219
2000	0.25	-1.00	0.00	0.00	0.00	CLT	0.0161	0.0030	0.0288	0.0100	0.0170	-0.0016	0.0220	-0.0479
						CLTcov	0.0078	-0.0141	0.0190	0.0140	0.0333	-0.0122	-0.0070	-0.0347
2000	0.25	-1.00	0.50	0.50	0.50	CLT	-0.0329	0.0514	0.0279	0.0332	-0.0187	-0.0108	-0.0262	0.0038
						CLTcov	-0.0429	-0.0033	0.0027	0.0224	-0.0156	-0.0196	-0.0209	-0.0220
2000	0.25	0.00	0.00	0.00	0.00	CLT	-0.0451	-0.0407	-0.0357	-0.0290	0.0114	-0.0230	-0.0165	0.0182
						CLTcov	-0.0366	-0.0234	-0.0325	-0.0373	0.0014	-0.0170	-0.0344	-0.0196
2000	0.25	0.00	0.50	0.50	0.50	CLT	-0.0046	-0.0059	-0.0128	0.0087	0.0190	0.0073	0.0142	-0.0037
						CLTcov	-0.0148	-0.0065	-0.0291	0.0013	0.0062	-0.0284	0.0085	-0.0386
2000	0.75	-1.00	0.00	0.00	0.00	CLT	0.0024	0.0021	0.0068	-0.0251	0.0266	0.0515	0.0029	0.0582
						CLTcov	0.0204	0.0175	-0.0181	-0.0421	0.0309	0.0300	-0.0394	0.0143
2000	0.75	-1.00	0.50	0.50	0.50	CLT	-0.0348	0.0241	-0.0131	-0.0322	-0.0441	-0.0180	-0.0021	0.0112
						CLTcov	-0.0406	0.0238	-0.0166	-0.0417	-0.0114	-0.0151	-0.0228	-0.0054
2000	0.75	0.00	0.00	0.00	0.00	CLT	-0.0081	-0.0157	-0.0091	-0.0034	-0.0484	-0.0401	-0.0219	-0.0615
						CLTcov	-0.0155	-0.0247	-0.0043	0.0001	-0.0204	-0.0536	-0.0292	-0.0614
2000	0.75	0.00	0.50	0.50	0.50	CLT	0.0172	0.0062	-0.0070	-0.0122	0.0062	-0.0334	-0.0240	-0.0060
						CLTcov	0.0044	0.0017	-0.0393	-0.0592	-0.0115	-0.0518	-0.0011	-0.0138

2 Data

The data are derived from the integration of the following five data sources:

(1) INVALSI administrative data collected in 2015 to assess the students' performance in the 5th grade in Italian and Mathematics. The Italian test includes two sections (Reading Comprehension and Grammar), whereas the Mathematics test consists of 27 items covering four main content domains (Numbers, Shapes and Figures, Algebra, Data, and Previsions). These tests are graded at the national level by advisers rather than by school teachers;

(2) INVALSI data collected in 2015 according to a questionnaire submitted to the students of all Italian schools to gather information on the NCSs. In addition, there is bullying acted and bullying right away that include words, actions, and images that hurt, humiliate, or socially exclude someone or that lower someone (inner stability);

(3) INVALSI administrative data collected in 2018 to assess the performance of the students in the 8th grade in Italian and Mathematics through tests having the same features as the data illustrated at point (1) above;

(4) PAT survey carried out in 2018 to collect demographics and socio-economic students' conditions, such as the parental socio-economic status related to the international socio-economic index named ESCS (index of economic, social and cultural status) defined within the Programme for International Student Assessment (OECD, 2015);

(5) PAT survey carried out in 2015 and 2018 concerning the features of the educational programs implemented by the schools (e.g., classes and teaching methodologies).

In the following, we provide additional details on the definition of non-cognitive skills (Table 6), the observed variables used in the application illustrating the proposal in Sections 5 and 6 of the paper (Table 7), and their summary statistics (Table 8). We also show the results of the test to compare the average achievement scores of participating and non-participating schools (Table 9). Finally, Table 10 reports additional descriptive statistics of the available data and the results of the t-test to compare student achievement scores at the baseline.

Table 6: *Description of the non-cognitive skills*

Definition	Description
Agreeableness	Collaboration and peaceful relationships
Extraversion	Attention to improve the school environment
Locus of control	Incentive to a personal position in making choices
Motivation	Development of affective dimensions linked to motivations
Openness to experience	Participation in the life of the local area
Conscientiousness	Identification of strengths and weaknesses
Emotional stability	Overcoming tense situations in daily life

3 Additional results of the application

3.1 Results of the proposed causal latent transition model

Figure 1 depicts the values of the BIC index reported in Table 1 of the paper, where we observe the increase in the BIC index for the models with more than two latent states.

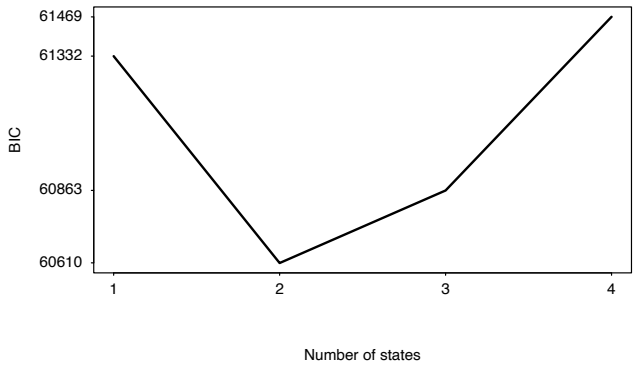


Figure 1: *Values of the BIC index of the CLT model for an increasing number of states*

Table 7: *Description of the variables available across years 2015 and 2018*

Responses	Description
Italian and Math scores at 5th grade†	Observed at the age of 10 years
Italian and Math scores at 8th grade†	Observed at the age of 14 years
Covariate	**Type (% of missing values)**
School motivations*	Continuous, standardized
Quality of class relations*	Continuous, standardized
External support for student autonomy*	Continuous, standardized
Well-being at school*	Continuous, standardized
Discomfort at school*	Continuous, standardized
Bullying acted*	Continuous, standardized
Bullying right away*	Continuous, standardized
Parents' ESCS index	Continuous, standardized (26%)
Employment status of father/mother**	Continuous (27%, 30%)
Italian nationality of the father	Binary (6%)
Gender	Binary (4%)

† *Continuous variable with higher scores indicating better results.* * *Variables collected in 2015.* ** *Values ranging from 0 to 7 with higher values indicating high employment level. The categories are ordered as follows: unemployed, house-maker, retired laborer, self-employed, professional, teacher, soldier, manager, professor, entrepreneur.*

Table 8: *Descriptive statistics of the INVALSI achievement scores at the 5th and 8th grade in Italian and Mathematics*

Achievement score	Italian		Mathematics	
	5th grade	8th grade	5th grade	8th grade
Min	80.238	81.604	111.141	71.691
q_1	185.259	188.654	184.784	191.454
Median	204.369	208.177	213.796	216.040
Mean	207.183	210.516	213.696	217.516
q_3	229.873	231.782	237.323	240.363
Max	343.456	353.584	364.753	372.015
s.d.	36.595	34.741	38.438	37.589

Table 9: *Average scores in Italian and Mathematics of pupils at the 5th grade attended schools participating and non-participating to the educational PAT programs, results of the usual t-test to compare schools*

	Italian		Mathematics	
Participating schools, mean (s.d.)	210.508	(6.623)	217.178	(7.508)
Non-participating schools, mean (s.d.)	208.468	(7.199)	214.791	(8.605)
Realized value of the t-statistics (p-value)	1.013	(0.237)	1.186	(0.196)

Table 10: *Descriptive statistics of the available data: overall, and treated and nontreated average values for the outcomes and covariates; relative frequencies for the binary covariates. Results of the usual t-test to compare the scores treated and nontreated students at the baseline ($t = 0$, 5th grade) for Italian and Mathematics*

Responses	Average	s.d.	Av.treated # 845	Av.nontreated # 716
Achievement score in Italian ($t = -1$)	196.900	34.212	197.588	196.677
Achievement score in Italian ($t = 0$)	207.183	36.595	207.586	206.708
Achievement score in Italian ($t = 1$)	210.516	34.741	210.636	210.375
Realized value of the t-statistics (p-value)			0.472	(0.637)
Achievement score in Mathematics ($t = -1$)	198.600	40.069	197.650	198.649
Achievement score in Mathematics ($t = 0$)	213.696	38.437	213.079	214.424
Achievement score in Mathematics ($t = 1$)	217.516	37.589	218.139	216.781
Realized value of the t-statistic (p-value)			-0.687	(0.492)
Continuous covariates	Average	s.d.	Av.treated	Av.nontreated
School motivation	-0.356	0.878	-0.262	-0.467
Parents' ESCS index	0.102	0.694	0.101	0.104
Quality of class relations	-0.001	0.512	0.011	-0.014
External support for student autonomy	0.000	0.529	0.017	-0.021
Well-being at school	-0.001	0.517	0.020	-0.025
Discomfort at school	0.001	0.700	0.024	-0.026
Bullying acted	0.000	0.482	0.002	-0.001
Bullying right away	0.000	0.675	0.008	-0.009
Employment status of father	2.579	2.116	2.786	2.335
Employment status of mother	2.301	2.149	2.488	2.081
Binary covariates	Proportion		Pr.treated	Pr.nontreated
Female	0.479		0.293	0.227
Italian nationality of the father	0.810		0.414	0.396

3.2 Results of the univariate causal latent transition model

In Table 11 we show the results of the estimated average treatment effects on the treated (ATETs) for the transition from the second to the first cluster and from the first to the second cluster of the univariate CLT model with k=2 states. These results confirm those obtained with the multivariate CTL model reported in Table 7 of the paper.

Table 11: *Estimates of the logit regression parameters of the treatment and standard errors (s.e.) for the transition probabilities under the univariate CLT model with k=2 states estimated separately for Italian and Mathematics: first column (Effect 1) from the first to the second cluster, second column (Effect 2) from the second to the first cluster (**significant at 1%)*

Scores	Effect 1	s.e.	Effect 2	s.e.
Italian	0.729	1.557	-3.935**	1.557
Mathematics	0.358	0.741	-4.229**	1.651

3.3 Results of the robust DiD model

In the following we show the average treatment effect (ATET) estimated with the locally efficient doubly robust difference-in-differences (DiD) estimator defined in Equation (3.1) of the paper Sant'Anna and Zhao (2020). Table 12 shows effects for the test results in Italian and Mathematics. We observe that these results are similar to those obtained with the standard DiD estimator presented in Tables 8 and 9 of the paper for Italian achievement score. For Mathematics, we get slightly different results from those presented in Table 9 of the paper since the effect is significant for Models 1 and 2 and not for Model 3, referred to students performing below the median score.

3.4 Additional analyses

To evaluate the consistency of the assumption related to the conditional Gaussian distribution of the outcomes, we compared the empirical distribution of each outcome once each

Table 12: *DiD estimator of the ATET for Italian with locally efficient doubly robust estimator proposed by Sant'Anna and Zhao (2020). Estimated effects of models for Italian, as in Equation (13) (top panel) and as in Equation (14) (bottom panel) of the paper, estimated for the overall students (Model 1), for students performing above the median score in grade 5th (Model 2) and for students performing below the median score in grade 5th (Model 3) (*significant at 5%, **significant at 1%)*

Equation (13)	Model 1	Model 2	Model 3
Italian	0.272	-0.036	2.431
Mathematics	4.052*	6.838**	1.867

Equation (14)	Model 1	Model 2	Model 3
Italian	0.846	0.410	1.603
Mathematics	3.534*	6.605**	1.326

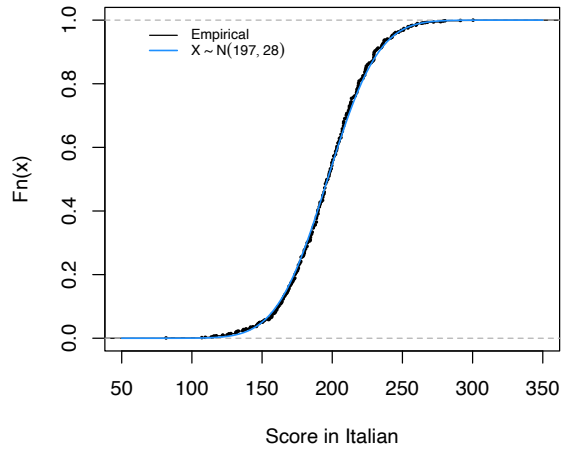
student has been assigned to a cluster according to the maximum-a-posteriori probability for each time occasion. This procedure is also known as local decoding (Bartolucci et al., 2013). Figures 2 and 3 show the conditional empirical distributions of each cluster for scores in Italian and Mathematics compared with the theoretical ones related to the first and second time occasions, respectively.

References

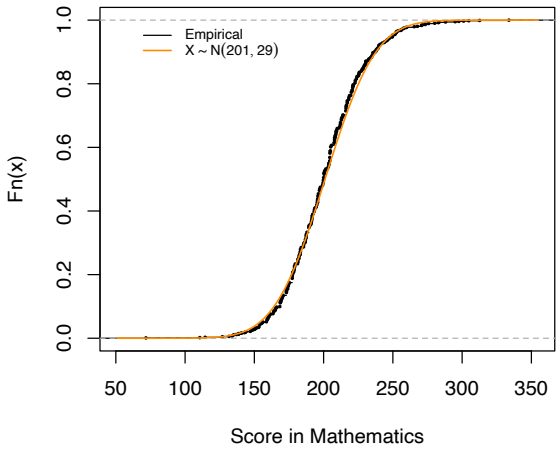
Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC, Boca Raton, FL.

OECD (2015). *Skills for Social Progress: The Power of Social and Emotional skills*. OECD Publishing.

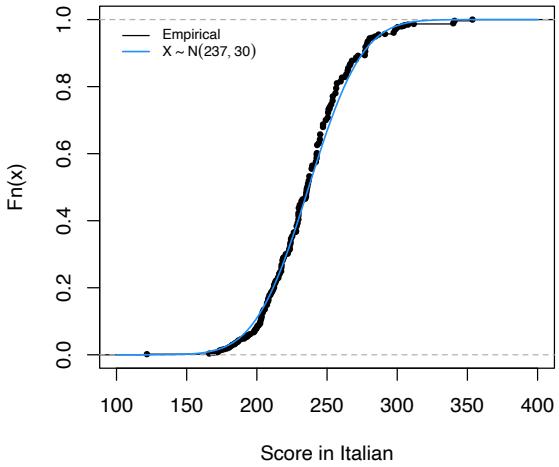
Sant'Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219:101–122.



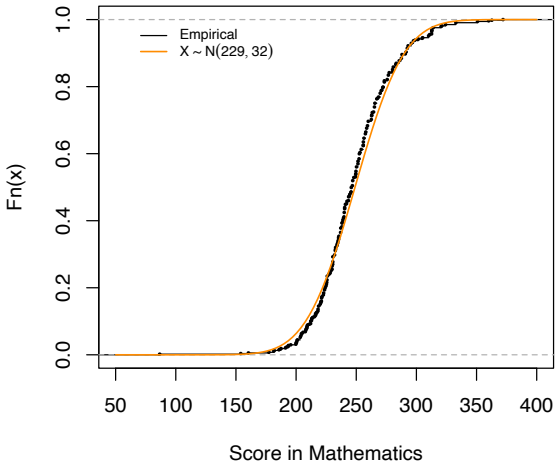
(a) Italian, cluster 1, 5th grade



(b) Mathematics, cluster 1, 5th grade

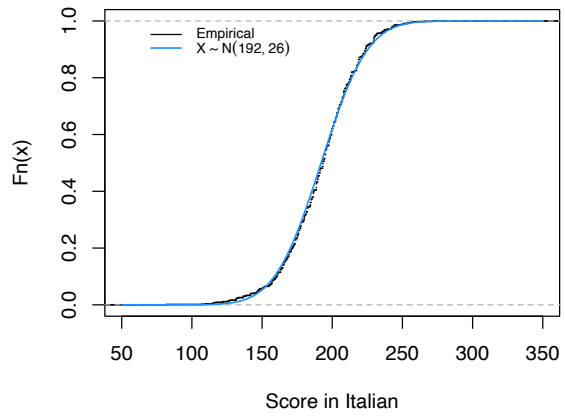


(c) Italian, cluster 2, 5th grade

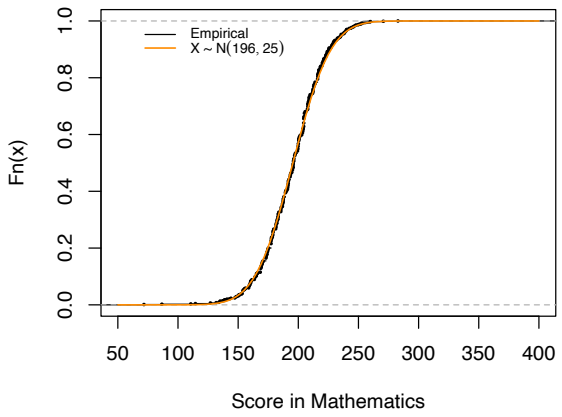


(d) Mathematics, cluster 2, 5th grade

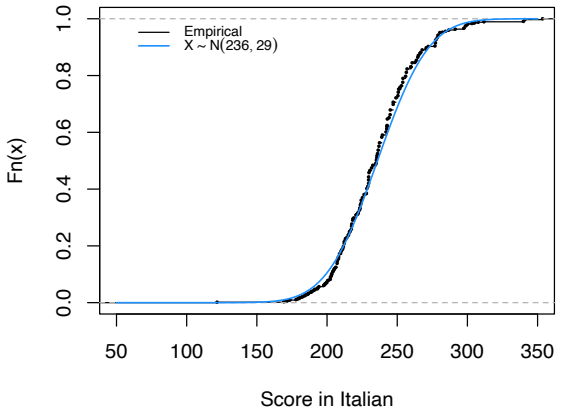
Figure 2: *Empirical and theoretical conditional cumulative distribution function for Italian and Mathematics obtained once the students have been assigned to a cluster on the basis of the maximum-a-posteriori rule at each time occasion*



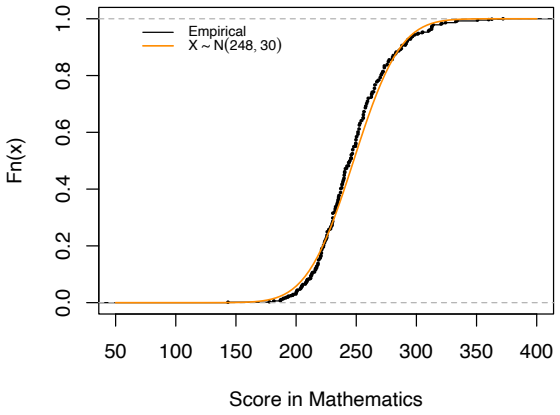
(a) Italian, cluster 1, 8th grade



(b) Mathematics, cluster 1, 8th grade



(c) Italian, cluster 2, 8th grade



(d) Mathematics, cluster 2, 8th grade

Figure 3: *Empirical and theoretical conditional cumulative distribution function for Italian and Mathematics obtained once the students have been assigned to a cluster on the basis of the maximum-a-posteriori rule at each time occasion*