DIRECT CAUSAL EFFECTS IN EDUCATION TRANSMISSION

VALENTINO DARDANONI, ANTONIO FORCINA, AND SALVATORE MODICA

Abstract. We exploit recent advances in latent class analysis to model a child's educational achievement as dependent on parents' education and the child's unobservable endowment. By simultaneously detecting and controlling for the child's unobservable endowment, we provide an estimate of the direct causal effect of parents' education on that of their child, in contrast to the total causal effect which requires controlling for the parents' unobservable endowment (a more problematic task). We apply our methodology to the NCDS dataset, a cohort study concerning English families in the nineteen-seventies. By looking separately at sons and daughters subsamples, we find that the direct effect we measure is significant in father-son relationships.

JEL Classification Numbers I21, C35

Keywords Education Transmission, Finite Mixtures, Direct Causal Effect, Likelihood Inference.

1. Introduction

This paper is an attempt to measure the *direct causal effect* of parents' education on children's educational achievement by conditioning on the children's unobserved endowment which, as explained below and in the appendix, may act as a confounder. Such conditioning is made possible by recent advances in latent class analysis (see for example Huang–Bandeen-Roche [26] and Bartolucci–Forcina [5]) which allow for modelling a causal diagram involving unobservable confounders, with only two restrictions: (i) unobservable variables must be discrete, (ii) the resulting parametric model must be identifiable. The implications of the first restriction, which are examined in more detail below, are not very demanding. On

Date: August 21, 2008.

the other hand, any meaningful model must be identifiable and we provide convincing evidence that our model is identifiable (see the end of section 3 and the appendix).

Estimation of the *total* effect of parents' education after controlling for their unobservable endowments has been the object of recent research initiated by Behrman and Rosenzweig [7]. Behrman and Rosenzweig arrive at the striking result that mothers' education has no effect on children's after controlling for parents' endowments by taking differences on MZ twin parents; in two important follow-ups of Behrman–Rosenzweig [7], Plug [31] using adoptees confirms the finding that only the father's education has a positive impact on the child's, while Black–Devereux–Salvanes [9], using reforms in municipal compulsory schooling laws as instruments, find almost no causal link between parents' and children's education.[1] The difficulties in controlling for parents' unobservable endowments are illustrated in the survey by Holmlund–Lindahl–Plug [25], where these three different methods (namely use of twins, adoptees and schooling laws instruments) are applied to a single data set, and it is shown that the three approaches produce results which are in conflict with each other.[2] The fact that estimates of the total causal effect may be quite sensitive to the key assumptions made in the process is compounded with the problem that separate estimation of fathers' and mothers' effects requires control for assortative mating.

To clarify the relation between direct and total effects consider the causal relation between parents' and children's schooling achievements represented by the

[1] On Behrman–Rosenzweig [7] see also the critical Comment [1] by Antonovics and Goldberger (and the authors' Reply [8]).

[2] " . . . using twin parents gives us positive intergenerational schooling coefficients for fathers, but a small or no effect for mothers . . . using foreign-born adoptees to identify the causal effect of parents education on childs education come out as relatively small . . . the IV strategy on the other hand, indicates that it is only the mothers education that is important, and that the effect for mothers is relatively large." [25], p.55.

following system of equations:

$$\begin{aligned} S^c &= f(S^f, S^m, \boldsymbol{Q}^c, \epsilon^c) && (1) \\ \boldsymbol{Q}^c &= \boldsymbol{g}(S^f, S^m, R^f, R^m, \boldsymbol{Q}^f, \boldsymbol{Q}^m, \eta^c), && (2) \end{aligned}$$

where S^c, S^f and S^m denote, respectively, child's, father's and mother's schooling; $\boldsymbol{Q}^c$,$\boldsymbol{Q}^f$ and $\boldsymbol{Q}^m$ are the vectors of unobservable child's, father's and mothers' endowments; R^f and R^m are father's and mother's child-rearing abilities; and ϵ^c, η^c are uncorrelated disturbances. The first equation says that a person's education depends on her own endowment and her parents' education. The effect of S^f and S^m on S^c, *given* the child's own endowment is what we call *direct effect*; the meaning that one can assign to such direct effects if detected is discussed below. The second equation can be considered a typical nature–nurture relation. Upon substitution, one obtains the standard reduced form equation (compare equation (2) in Behrman and Rosenzweig [7] or equation (8) in Holmlund *et al.* [25]):

$$S^c = f(S^f, S^m, R^f, R^m, \boldsymbol{Q}^f, \boldsymbol{Q}^m, \theta^c). \qquad (3)$$

The model presented above implies that the effect of parents's schooling on S^c is both *direct*, by equation (1), and *indirect* acting first on the vector of unobservable endowments $\boldsymbol{Q}^c$, as stated in (2). If one could control properly for parent's endowments, equation (3) would allow an estimate of the total causal effect of parent's schooling. On the other hand, because (1) and (2) imply that S^c is independent of $\boldsymbol{Q}^f$, $\boldsymbol{Q}^m$ conditionally on S^f, S^m and $\boldsymbol{Q}^c$, controlling for $\boldsymbol{Q}^c$ alone allows estimation of the direct causal effect of parents' schooling, which is the object of the present paper. In the appendix (section 6.1) we provide a formal decomposition of the total effect into the direct and indirect effects and discuss why one cannot obtain an unbiased estimate of the total causal effect without controlling for parents' unobservable endowments. The book by Pearl [30] contains an exhaustive analysis on estimation of causal effects; Chalak and White [14] discuss many examples of applications of causal analysis in econometrics.

To analyze the direct causal effect of parents' schooling we consider as dependent variable, rather than years of schooling, an indicator of schooling *attainment* in terms of achievement of a significant educational certification, since this is more likely to reflect the value assigned to formal education by the students and their parents. We use the English NCDS dataset, where such a certification is represented by the O-Level exams (details in section 2), and control for children's unobservable endowment at the time of the attainment of the scholastic certification (in our case 16 years of age).[3] This is achieved by exploiting the features of the dataset, which contains information on a rich set of variables concerning different kinds of abilities measured from very early age, and we identify the unobserved children's schooling endowment vector $\boldsymbol{Q}^c$ by a finite mixture model. Finite mixture models allow both the marginal distribution of the latent and the conditional distribution of the responses to be unconstrained, the only limitation being that the number of support points is finite, which means that there is a finite collection of distinct qualitative *types*; because of this, in the sequel we rewrite $\boldsymbol{Q}^c$ as a discrete random variable U taking values in $\{1, \ldots, m\}$. The restrictions implied by the discrete qualitative nature of U is that latent types which are not sufficiently distinct will be clumped together. Thus, unless additional structure is imposed on the latent classes, they can capture the multidimensional nature of endowments often stressed, for instance, in the labor market literature (see e.g. Heckman-Stixrud-Urzùa [23]).

The natural interpretation of the direct effect we look at emerges by asking why, among subjects with a given level of schooling endowment, those with more

[3] In particular, the analysis of *evolution* of endowments from early age is outside the scope of the paper. This important line of research is actively pursued by Heckman and associates, cfr. e.g. [16, 17, 23], who also arrive at evaluating dynamic complementarities of early and late interventions in the formation of skills.

educated parents should attain higher educational levels; possible explanations include, for example, the intergenerational transmission of education-dependent labor markets skills, better information on the value of education, or simply greater parental pressure reflecting social norms. In any given social context, note that this kind of influence (which may be called a *role effect*) is typically gender dependent, in the sense that mothers and fathers may have different effects on daughters and sons.[4] The empirical findings we report confirm the presence of role effects in our sample: given the child's schooling ability, more educated parents bring their offsprings to a greater level of education. The nature of the effect we estimate appears more neatly by examining its gender dependence: when the sons and daughters samples are analyzed separately, it is seen that fathers' education affects *only* that of their sons; mothers' education has a weak (and not significant) effect, only on daughters. It may be worth noting that this strong father–son link may well be confined to the social context to which our data refers. We discuss these findings in the conclusions after we present estimation results.

The paper is organized as follows. The data are described in section 2; we then derive the model to be estimated (section 3) and report the results of estimation (section 4). Section 5 contains some concluding remarks. In the appendix we explain the details of the identification and estimation of our model using a likelihood inference approach.

2. The Data Structure

In the British educational system, students at the age of 16 take the so called O-Level exams on a set of chosen topics. If a student has reached a minimum standard

[4] In studies of educational attainment, early mention of the possibility of interclass differences in educational choice at given levels of academic performance is found in Boudon [11], who called this a 'secondary effect'. Erikson *et al.* [20] confirm the presence of this effect by counterfactual analysis.

in terms of quantity of subjects taken and grades obtained, she is awarded an O-Level certification and allowed, if she wishes, to access the next level of education (the so-called A-Level).

We use data from the National Child Development Survey (NCDS). This data set is a UK cohort study targeting all the population born in the UK between the 3rd to the 9th of March 1958. Individuals were surveyed at different stages of their life and information on their schooling results and their background was collected. Our main dependent variable is the binary variable OL which takes value 1 iff the subject has obtained the necessary formal O-Level qualifications; in our sample about fifty percent of the subjects achieve them.[5] [6]

In the NCDS subjects are tested at the age of 7, 11 and 16 for mathematics, reading and general cognitive skills, and at the age of 7 and 11 information on non-cognitive skills is also collected; we use the results of these tests for identification of the unobservable endowment U. In particular, we first replace, at each age, the original maths and reading scores with the principal component (in all cases this explains no less than 90% of the total variance). For non-cognitive skills (available at ages 7 and 11) analogous factor analysis yields two factors, where the first factor concerns ability to relate to other individuals, while the second captures emotional problems. We then extract six binary response variables by appropriately dichotomizing these cognitive and noncognitive variables; we call them EM, LM, ER, LR, PNC, SNC, which are meant to capture early (7 and 11) and late (16) math and reading endowments, and (early) personal and social noncognitive skills.[7] Parents' schooling is defined as the age at which they left

[5] Since attainment of O-Level certification is the first significant schooling continuation decision made by UK students, considering this as the dependent variable avoids the dynamic selection bias problem pinned down by Cameron-Heckman [12].

[6] The same data set is extensively used, among others, by Blundell–Dearden–Sianesi [10], who study the effect of education on earnings.

[7] Late binary variables take value 1 if the corresponding variable is above the sample median; early binary variables take value 1 if the corresponding variables are greater than the median both at 7 and 11.

school; fathers' and mothers' schooling are collected into the vector $\boldsymbol{s}$. Regarding other family background variables, the NCDS contains also information on parents' interest in their child's education, as reported by teachers separately for mothers and fathers; these can be considered a proxies for the child-rearing abilities R^f, R^m. Data on parents' interest are originally classified into 5 distinct categories (overconcerned, very interested, shows some interest, little interest, can't say); we have created two parent's interest binary variables which take value one if the parent is in one of the first three categories at all ages considered (7, 11 and 16). The two parents' interest variables are collected in the vector $\boldsymbol{z}$; the vector of all four family background characteristics is denoted by $\boldsymbol{x} = (\boldsymbol{s}, \boldsymbol{z})$.[8]

From the NCDS we selected all subjects for whom we had information on OL, test scores and $\boldsymbol{x}$; the resulting sample is made of 5195 individuals, 2627 sons and 2568 daughters. Summary statistics on the data used are reported in Table 1 below. A complete description of the data is available at

`http://www.esds.ac.uk/longitudinal/access/ncds`.

	Avg daughters	Avg sons	short name	type
O-Level	0.5674	0.4943	OL	dummy
Early Math	0.3910	0.4201	EM	dummy
Late Math	0.5047	0.5869	LM	dummy
Early Reading	0.4276	0.3730	ER	dummy
Late Reading	0.4938	0.5321	LR	dummy
Personal NC Skills	0.4007	0.2636	PNC	dummy
Social NC Skills	0.3551	0.2666	SNC	dummy
Father schooling	14.9743	14.9351	fs	numerical
Mother schooling	15.0195	14.9518	ms	numerical
Father interest	0.4178	0.4227	fi	dummy
Mother interest	0.3812	0.3856	mi	dummy

Table 1. Data

[8] We also considered parents' age, but dropped it from the analysis because it was never found significant.

3. Unobserved endowments and finite mixture models

In this section we present the finite-mixture model we estimate. We briefly recall the setting of classical latent class analysis, then turn to the extensions which our set-up incorporates.

3.1. **Notations and introductory remarks.**

The finite-mixture approach is used in many branches of statistics such as biometrics and psychometrics (see e.g. [29, 33]). An early use in economics is in Heckman and Singer [24]; recently, it has been mainly used in dynamic models, with several applications to scholastic achievements (notably Keane and Wolpin [27, 28]). Arcidiacono and Jones [2] contains an ample discussion of this literature. We introduce here the basic notation.

Consider a set of k binary response variables Y_j, $j = 1, \dots, k$, taking values $y_j \in \{0, 1\}$, where, as usual, lower case letters denote observed values of the corresponding capital-letter random variables. A given response configuration will be denoted by the column vectors $\boldsymbol{y} = (y_1, \dots, y_k)'$ and $q_{\boldsymbol{y}}$ will denote the probability of a given response configuration. Now order the response patterns lexicographically, with elements on the right running faster from 0 to 1. A probability distribution on the set of the 2^k distinct response configurations will be represented by the vector $\boldsymbol{q}$ having elements $q_{\boldsymbol{y}}$, ordered as above; this vector belongs to the simplex Δ_{2^k}.

Now, since it is often inconvenient to work directly with the probability vector $\boldsymbol{q}$, one typically uses a set of *parameters* which are simply functions from Δ_{2^k} to $\mathbb{R}^{2^k-1}$. To parameterize multivariate distributions we use marginal and conditional logits; below we recall the basic ideas for the reader's convenience. Let $\boldsymbol{q} \in \Delta_4$ denote the distribution of two binary response variables Y_1 and Y_2. One can describe $\boldsymbol{q}$ by an appropriate choice of 3 free parameters; for example, $\boldsymbol{q}$ could be defined by two parameters describing the univariate marginals and a parameter describing their association such as $\Pr(Y_i = 1)$, $i = 1, 2$ and $\Pr(Y_1 = 1, Y_2 = 1)$. However, this parameterization is not unique. Another invertible mapping from Δ_4 to $\mathbb{R}^3$ is obtained using the *logit* $\lambda_{Y_1} \equiv \ln[\Pr(Y_1 = 1)/\Pr(Y_1 = 0)]$ to describe the marginal distribution of Y_1 and the two logits $\lambda_{Y_2|Y_1=0}$ and $\lambda_{Y_2|Y_1=1}$ to represent the conditional distribution of Y_2. These are equivalent alternative parameterization of $\boldsymbol{q}$ since they convey all relevant information on the joint distribution of Y_1, Y_2 (see e.g. Bartolucci–Colombi–Forcina [3] for a general recursive definition of logit parameters and a discussion on their invertibility properties).

Clearly, assuming that some interaction parameters are zero or that some conditional logits are equal, implies that $\boldsymbol{q}$ belongs to a subset of Δ_{2^k}; for example, if we assume that Y_1 and Y_2 are independent, then any $\boldsymbol{q} \in \Delta_4$ which satisfies the above relation can be uniquely described by two parameters such as $\Pr(Y_1 = 1)$, $\Pr(Y_2 = 1)$ or equivalently, e.g., λ_{Y_1} and λ_{Y_2}.

3.2. **Classical latent class analysis.** Given observable binary responses $(Y_1, \ldots, Y_k)$, classical latent class analysis (e.g. Goodman [22]) tries to identify a discrete random variable U taking values in $\{1, \ldots, m\}$ such that

$$\Pr(Y_1 = y_1, \ldots, Y_k = y_k) = \\ \sum\nolimits_u \Pr(U = u) \Pr(Y_1 = y_1 \mid U = u) \cdot \ldots \ldots \cdot \Pr(Y_k = y_k \mid U = u); \quad (4)$$

that is, the unobservable latent variable U makes observed responses conditionally independent. Clearly this assumption (which is known in the latent class literature as *local independence*) restricts the dimension of the probability space of $(U, Y_1, \ldots, Y_k) \equiv (U, \boldsymbol{Y})$ from $m \cdot 2^k - 1$ to $(m - 1) + m \cdot k$. Indeed, any $\boldsymbol{p} \in \Delta_{m \cdot 2^k}$ which satisfies relation (4) is uniquely determined by the following parameters: $\Pr(U = u)$ for $u = 1, \cdots, m - 1$, and $\Pr(Y_i = 1 \mid U = u)$ for $i = 1, \ldots, k$, $u = 1, \ldots, m$. Notice however that, since U is not observed, the dimension of the space of the observed responses is only $2^k - 1$, and this restricts the number of classes of U which can be identified, since $(m-1) + m \cdot k$ must be less than or equal to $2^k - 1$.

3.3. **The model estimated in this paper.** To identify the unobservable child's endowment $U \in \{1, \dots, m\}$ one can extract information from response vectors which act as *multiple indicators* of the latent variable U. As indicators we shall use dichotomized observations on maths and reading test scores and non cognitive skills as defined above. To be more specific note that in the case at hand the unobservable residual heterogeneity affecting a given educational attainment (the English O-Level certification in our case) can be seen as the result of two different factors: early cognitive and noncognitive endowments and relevant knowledge acquired through learning. Since this is the individual heterogeneity which our latent variable U should capture, we shall have it *jointly* identified, besides OL itself, by early indicators of innate mathematical and reading comprehension, by indicators of the level of acquired mathematical and reading knowledge at the time OL exams are taken, and by personal and social noncognitive skills. So in our case $k = 7$ and $\boldsymbol{Y} = [OL, EM, LM, ER, LR, NCP, NCS]$. If a random variable U which satisfies (4) could be identified, such variable would capture the underlying unobserved endowment, since this would imply that, conditionally on U, not only knowledge of any test score would be irrelevant for predicting the OL results, but also that, for instance, knowledge of EM is irrelevant for predicting LM.

However classical latent class models, in particular the assumption underlying equation (4), are restrictive. The first point to notice is that it seems plausible that endowments and responses are themselves affected by family background characteristics. Thus, a first extension of the classical model is to allow the distribution of the response vector $(U, \boldsymbol{Y})$ to depend on observable covariates, which we recall are denoted by $\boldsymbol{x} = (\boldsymbol{s}, \boldsymbol{z})$.[9] Furthermore, it seems plausible that test results taken on the same subject matter may still be dependent even after conditioning on U. We then weaken (4) by considering dependence on covariates and allowing some conditional dependencies:[10]

Assumption 1.

$$\Pr(\boldsymbol{y} \mid \boldsymbol{x}) = \sum\nolimits_u \Pr(u \mid \boldsymbol{x}) \Pr(ol \mid u, \boldsymbol{s}) \Pr(\boldsymbol{m} \mid u) \Pr(\boldsymbol{r} \mid u) \Pr(ncp \mid u) \Pr(ncs \mid u) \tag{5}$$

where $\boldsymbol{m} = (em, lm)$ *and* $\boldsymbol{r} = (er, lr)$*.*

Assumption 1 makes it explicit the way in which family background influences observed responses:

- writing $\Pr(\boldsymbol{r}|u)$, $\Pr(\boldsymbol{m}|u)$, $\Pr(ncp \mid u)$ and $\Pr(ncs \mid u)$ as independent of $\boldsymbol{x}$ implies that we are using a kind of *absolute performance* in cognitive and noncognitive skills without correcting for different backgrounds to identify U; however, while math and reading test scores are assumed conditionally independent, residual association is allowed within $\boldsymbol{m}$ and $\boldsymbol{r}$;
- on the other hand, we allow the marginal distribution of U to depend on family background variables;
- we model the conditional distribution $\Pr(OL \mid U = u, \boldsymbol{x})$ to estimate the effect of a change in parents' schooling $\boldsymbol{s}$ on OL while controlling for U.

As it can be seen from equation (5), the distribution of $\boldsymbol{Y}$ conditional on U and $\boldsymbol{x}$ is decomposed into five conditionally independent blocks, namely OL results, math and reading test scores and personal and social non cognitive skills. Our

[9] Huang–Bandeen-Roche [26] explain how a finite mixture model can be identified and estimated in the presence of continuous and discrete covariates, under the local independence assumption.

[10] A latent class model where both the responses and the latent variable are allowed to depend on covariates, and residual association is allowed on responses, is described in Bartolucci–Forcina [5] in the context of capture/recapture models.

second assumption is that math and reading test scores follow a first order Markov recursive system:

Assumption 2.

$$\Pr(\boldsymbol{m} \mid u) = \Pr(em \mid u) \Pr(lm \mid em, u);$$

$$\Pr(\boldsymbol{r} \mid u) = \Pr(er \mid u) \Pr(lr \mid er, u).$$

We remark that, unlike in the analyses of Heckman and collaborators [16, 17, 23], the dynamics of the unobservable individual heterogeneity U are not analyzed in this paper. However, since our interest lies in capturing the unobservable schooling endowments at the time where OL exams are taken, formally U can be seen as a cross-classification of underlying abilities both over different cognitive and non cognitive skills and over times; what matters is that a sufficient number of types is used to capture the unobservable heterogeneity within a model which is identifiable.

Let now $\boldsymbol{p}(\boldsymbol{x}) \in \Delta_{m \cdot 2^7}$ denote the probability vector which describes the conditional distribution of $(U, \boldsymbol{Y} \mid \boldsymbol{x})$; and let

$$\Delta^o(\boldsymbol{x}) = \{\boldsymbol{p}(\boldsymbol{x}) \in \Delta_{m \cdot 2^7} \mid \boldsymbol{p}(\boldsymbol{x}) \text{ satisfies Assumptions (1)–(2)}\}$$

denote the set of $\boldsymbol{p}(\boldsymbol{x})$'s which can be decomposed according to Assumptions 1 and 2. The dimension of this parameter space is $v = (m-1) + m(5+2+2)$, since there are $m-1$ marginal weights for U and the conditional distribution of $Y|U = u$ require 2 logits each for LM and LR while only one is needed for the other 5 remaining responses.

Finally, let $\boldsymbol{\lambda}(\boldsymbol{x})$ denote the v-dimensional vector that collects the following logits:

$$\begin{aligned}\boldsymbol{\lambda}(\boldsymbol{x}) \;=\; & [\lambda^{2}_{U|\boldsymbol{x}}, \ldots, \lambda^{m}_{U|\boldsymbol{x}};\ \lambda_{EM|U=1}, \ldots, \lambda_{EM|U=m}; \\ & \lambda_{ER|U=1}, \ldots, \lambda_{ER|U=m};\ \lambda_{LM|U=1,EM=0}, \ldots, \lambda_{LM|U=m,EM=1}; \\ & \lambda_{LR|U=1,ER=0}, \ldots, \lambda_{LR|U=m,ER=1};\ \lambda_{NCP,U=1}, \ldots, \lambda_{NCP,U=m}; \\ & \lambda_{NCS,U=1}, \ldots, \lambda_{NCS,U=m};\ \lambda_{OL|U=1,\boldsymbol{s}}, \ldots, \lambda_{OL|U=m,\boldsymbol{s}}]'\end{aligned}$$

where $\lambda^{u}_{U|\boldsymbol{x}} = \ln\left[\Pr(U = u \mid \boldsymbol{x})/\Pr(U = u-1 \mid \boldsymbol{x})\right]$, $u = 2, \ldots, m$, denote the consecutive logits which are appropriate in this context given the multinomial nature of U.

Using recent theory of marginal modeling it can be shown that any conditional distribution $\boldsymbol{p}(\boldsymbol{x}) \in \Delta^{o}(\boldsymbol{x})$ can be conveniently parameterized in terms of $\boldsymbol{\lambda}(\boldsymbol{x})$ without imposing any parametric restrictions besides those implied by the Assumptions (1)–(2); the relation between $\boldsymbol{p}(\boldsymbol{x})$ and $\boldsymbol{\lambda}(\boldsymbol{x})$ is stated formally in the following:

Proposition. *For any $\boldsymbol{p}(\boldsymbol{x}) \in \Delta^{o}(\boldsymbol{x})$ with strictly positive elements, the mapping between $\boldsymbol{p}(\boldsymbol{x})$ and $\boldsymbol{\lambda}(\boldsymbol{x})$ is invertible and differentiable.*[11]

Proof. The result follows from the factorization in Assumptions (1)–(2) and the fact that the mapping between a univariate discrete distribution and the corresponding set of logits is a diffeomorphism. The result is also a special case of

[11] The mapping from $\boldsymbol{p}(\boldsymbol{x})$ to $\boldsymbol{\lambda}(\boldsymbol{x})$ can be written in explicit form by constructing an appropriate contrast matrix $\boldsymbol{C}$ (whose rows have elements summing to zero) and a marginalization matrix $\boldsymbol{M}$ (a matrix made of 1's and 0's) such that, for any $\boldsymbol{p}(\boldsymbol{x}) \in \Delta^{o}(\boldsymbol{x})$, we have

$$\boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{C}\ln(\boldsymbol{M}\boldsymbol{p}(\boldsymbol{x})).$$

In particular, $\boldsymbol{C}$ is a block diagonal matrix with elements equal to $(-1 \quad 1)$. For each block of $\boldsymbol{C}$ the matrix $\boldsymbol{M}$ has two rows of length $(m+1)\cdot 2^{7}$ which select the elements of $\boldsymbol{p}(\boldsymbol{x})$ that constitute the two events to be compared in the corresponding logit. Colombi and Forcina [15] show how these matrices can be constructed.

Theorem 1 in Bartolucci–Colombi–Forcina [3] who study the properties of a general class of marginal parameterizations which constitute *link functions*, that is re-parameterizations which are one to one and at least twice differentiable. □

The proposition above implies that the mapping from $\boldsymbol{p}(\boldsymbol{x}) \in \Delta^o(\boldsymbol{x})$ to $\boldsymbol{\lambda}(\boldsymbol{x})$ defines an invertible link function $\boldsymbol{h} : \mathbb{R}^v \mapsto \Delta^o(\boldsymbol{x})$ such that any $\boldsymbol{p}(\boldsymbol{x}) \in \Delta^o(\boldsymbol{x})$ can be written as $\boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{h}\big(\boldsymbol{\lambda}(\boldsymbol{x})\big)$. Thus, for each $\boldsymbol{x}$, $\boldsymbol{\lambda}(\boldsymbol{x})$ conveys all relevant information on $\boldsymbol{p}(\boldsymbol{x})$.

If the covariates were discrete with a very limited number of distinct configurations (strata), these could be coded with a corresponding number of dummy variables. This would be equivalent to fit a separate model to each stratum and the model could be called saturated (see e.g. Wooldridge [49] p.456 for terminology), since there would be no sharing of parameters across strata. On the other hand, if covariates are continuous or take on so many values that most strata contain only one subject, the approach just described is not viable. We then model $\boldsymbol{\lambda}(\boldsymbol{x})$ as a linear function of the covariates, and estimate the following multivariate regression system:

$$\begin{aligned}
&\lambda_{OL|u,\boldsymbol{s}} = \alpha_{OL}(u) + \boldsymbol{s}'\boldsymbol{\beta}_{OL}\\
&\lambda_{EM|u} = \alpha_{EM}(u)\\
&\lambda_{LM|u,em} = \alpha_{LM}(u) + \gamma_{LM}(u)\, em\\
&\lambda_{ER|u} = \alpha_{ER}(u)\\
&\lambda_{LR|u,er} = \alpha_{LR}(u) + \gamma_{LR}(u)\, er\\
&\lambda_{NCP|u} = \alpha_{NCP}(u)\\
&\lambda_{NCS|u} = \alpha_{NCS}(u)\\
&\lambda^u_{U|\boldsymbol{x}} = \alpha_U(u) + \boldsymbol{x}'\boldsymbol{\beta}_U(u),\ u = 2, \dots, m,
\end{aligned} \tag{6}$$

which can be written more compactly as

$$\boldsymbol{\lambda}(\boldsymbol{x}) = \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\psi} \tag{7}$$

where $\boldsymbol{B}(\boldsymbol{x})$ is a design matrix whose dependence on $\boldsymbol{x}$ reflects the effect of the covariates on the different elements of the joint distribution, and $\boldsymbol{\psi}$ is the vector which collects the model parameters α's, β's and γ's.

The standard approach to parameters' estimation in finite mixture models is the EM algorithm whose implementation to our context is described in the appendix. The basic idea of the EM algorithm is that, if the joint frequency table $(U, \boldsymbol{Y} \mid \boldsymbol{x})$ were known, maximum likelihood would be equivalent to estimation of a regression model within the multinomial distribution. At the E (expectation) step the unobservable frequency table is replaced by its expected value computed conditionally on the observed frequency table and the (possibly updated) estimates of the model parameters.

It has been shown (Dempster–Laird–Rubin [19]) that the algorithm converges to the maximum of the true likelihood. It is well known that the EM algorithm converges even if the model is not identified, a crucial issue for finite mixture models. The methods proposed by Forcina [21] have been applied to test that the model is locally identifiable for a wide range of parameter values. A brief formal discussion of the problem is contained in the appendix.

4. Results

We start by estimating model (6) under a different number of unobservable types. Maximum likelihood estimation is performed by an EM algorithm as described in the appendix; computations are based on a set of Matlab functions available upon request. The maximized log-likelihood $L(\hat{\boldsymbol{\psi}})$ and Schwartz's Bayesian Information Criterion $BIC(\hat{\boldsymbol{\psi}}) = -2L(\hat{\boldsymbol{\psi}}) + \ln(n)v$, where n denotes sample size and v is the number of parameters (which depends on the number of latent classes m), are given in Table 2 below. Since $BIC(\hat{\boldsymbol{\psi}})$ is lowest with three latent classes in both samples, these results seem to indicate that three latent classes are adequate to represent unobserved heterogeneity.

In addition, a comparison of the estimated coefficients for parents schooling in the *OL* equation reveals that, while with two latent classes the estimates are sensibly greater than the corresponding ones with three and four classes, the differences between estimates based on three and four classes are negligeable.[12] Next, for parsimony and sharpness of parameters' estimation, we tested the restriction that the slopes $\gamma_{EM}(u)$ and $\gamma_{ER}(u)$, which give the recursive structure to test results, do not depend on U. Since U has three levels, this imposes a total of $2 \times 2 = 4$ linear constraints. The restriction is not rejected by a standard LR test (the test statistics are respectively equal to 2.632 and 3.574 in the daughters and sons subsamples).

		Daughters		Sons	
latent cl.	param.	$L(\hat{\boldsymbol{\psi}})$	$BIC(\hat{\boldsymbol{\psi}})$	$L(\hat{\boldsymbol{\psi}})$	$BIC(\hat{\boldsymbol{\psi}})$
2	23	-9770.4	19721	-9816.8	19807
3	35	-9665.7	19606	-9679.9	19636
4	47	-9647.9	19665	-9647.9	19672

TABLE 2. Maximized log-likelihood and BIC

We next enquire whether the unobservable endowments are indeed multidimensional by testing the *monotonicity* hypothesis (see for example Bartolucci and Forcina [4], section 2), which requires that, after a suitable reordering of the latent classes, the conditional expectation of each response variable is an increasing function of the latent; thus, the monotonicity assumption is equivalent to assume that the model parameters satisfy a sistem of linear inequalities, and is also equivalent to assume the existence of a unique underlying unobservable ordered discrete variable representing endowments. With our estimates, not a single inequality is violated both in the sons and daughters subsamples; this implies that the LR test statistic for testing monotonicity against a unspecified alternative is equal to

[12] Estimated coefficients for two and four latent classes are not reported for the sake of brevity.

zero and thus the null cannot be not rejected. We then conclude that subjects with $U = 1$ are the least endowed in all cognitive and noncognitive tests, those in $U = 3$ are the most talented, with $U = 2$ individuals being somehow in the middle (see the α-coefficients in table 6). However, by looking at how much these coefficients increase when going from $U = 1$ to $U = 2$ and then to $U = 3$, we note that different groups of responses define, somehow, a different metric meaning that certain unobservable "types" are more distant according to some responses and closer relative to others.

4.1. **Parameters' estimation.** From estimation of system (6), whose results are reported in Table 6 in the appendix, the direct effects described in the introduction emerge.[13] Indeed, the coefficients of the parents' education variables in the *OL* equation (first two rows in the Table) show that after controlling for the child's schooling endowments, the father's education significantly helps his son's chance of achieving *OL* certification, and not his daughter's; and that on the other hand, mothers' education has no statistically significant effects.

The following facts about endowments seem also worthy of notice:

- The effect of being a high rather than low U type is rather substantial on all response variables, as can be seen by looking at the α-coefficients (recall that in the logit scale a change of value, say, from -1 to +1 implies a change in the probability of success from 27% to 73%). As one would expect, the effect is more pronounced for mathematical and reading abilities than for noncognitive skills.
- There is a strong positive association between child's endowment U and family background characteristics (as measured by the last equation of (6)), since having more educated or more concerned parents increases the odds of both being a type $U = 3$ rather then $U = 2$, and also a type $U = 2$

[13] To ease interpretation of the intercepts all covariates have been centered.

rather then $U = 1$. This is seen from the positive significant values of $\boldsymbol{\beta}_U$ for both $U = 3, 2$. However, as we discuss in the appendix (section 6.1), since we do not control for parents' endowments, these estimates of the indirect effect may be biased.

- Even after conditioning on U, there is still a strong positive correlation between test score results in reading taken at earlier and later stages, as it emerges from the significantly positive estimates of γ_{LR} in both subsamples and the fact that γ_{LM} is significant in the daughters subsample. This may be taken an an indication that certain subsets of responses require, in addition to the overall endowment captured by U, a certain amount of specific abilities.

4.2. **Pooled Sample.** For the sake of comparison we have estimated the previous model on the pooled sample of sons and daughters. As can be seen by comparing the corresponding lines of Tables 6 and 3 below, the pooled sample estimates are approximately equal to the average of the corresponding estimates for daughters and sons. They suggest a somewhat stronger role of fathers, but the coefficients are not significant. Thus considering sons and daughters subsamples separately is crucial for clarifying the nature of this asymmetry.

	$\boldsymbol{\beta}_{OL}$ coeff	se
father sch	0.0452	0.0380
mother sch	0.0253	0.0403

TABLE 3. Direct effects in pooled sample

4.3. **Direct effects.** We now translate the above estimates into a quantitative appraisal of the direct effect. We consider the effect on OL attainment of increasing a parent education by three years of schooling (which seems a reasonable measure of change of educational status), leaving unchanged the schooling of the other parent, starting from a situation where both parents have an average level of schooling (here denoted by μ^i, $i = f, m$).

The average effect of increasing a parent's education for a given level of child's endowment U can be calculated as:

$$\begin{aligned}\delta(u, S^i) &= \Pr(OL = 1 \mid S^j = \mu^j, S^i = \mu^i + 3, U = u) - \\ &\quad \Pr(OL = 1 \mid S^j = \mu^j, S^i = \mu^i, U = u) \\ &= \Lambda(a_{OL}(u) + 3\beta_{OL,S^i}) - \Lambda(a_{OL}(u)), \quad i, j = f, m, \quad u = 1, 2, 3,\end{aligned}$$

where $\Lambda(t) = \exp(t)/(1 + \exp(t))$ denotes the logit link function, and the second equation follows since we have centered father's and mother's schooling. These effects can be consistently estimated using the coefficients in the OL-equation; furthermore, using the estimated variance matrix of $(a_{OL}(u), \beta_{OL,S^i})$, an asymptotic standard error can be derived by application of the delta method. Their numerical values are in the following table:

	Daughters		Sons	
	$\delta(u, S^f)$	se	$\delta(u, S^f)$	se
U=3	-0.0059	0.0098	0.0263	0.0129
U=2	-0.0121	0.0380	0.0829	0.0392
U=1	-0.0040	0.0137	0.0253	0.0122
	$\delta(u, S^m)$	se	$\delta(u, S^m)$	se
U=3	0.0315	0.0248	-0.0068	0.0129
U=2	0.0555	0.0403	-0.0270	0.0442
U=1	0.0171	0.0116	-0.0096	0.0156

Table 4. Direct Effects of Father's and Mothers' Schooling

The table shows that the direct effect, measured as difference in probabilities of achievement, is rather substantial and statistically significant only for fathers on sons; notice also that direct effects are highly nonlinear in U.

4.4. **Proxying U.** As an alternative strategy to identify unobservable endowments, it would be reasonable to *proxy* U by the rich set of early cognitive and noncognitive test results. While this has the advantage that there is no need to dichotomize such variables, estimation results are subject to the usual assumptions which make use of proxy variables valid for causal analysis (see e.g. Wooldridge [34] p.63). A logit regression of OL on parents' schooling and interest, test score results at 7, 11 and 16 in reading and math and personal and social non cognitive skills at 7 and 11 gives the estimated coefficients for the direct effect of parents schooling reported in Table 5 below. As it can be seen by comparing the estimated coefficients with those of Tables 6 and 3, the simple use of proxies for U gives a picture of the direct effect of parents' schooling which is broadly comparable to our estimates; however, it seems that, at least in these samples, the simple proxy method tends to substantially overestimate the direct effects of interest, thus hinting at its inadequacy to fully control for children's unobservable endowments compared to our finite mixture model.

	Daughters		Sons		All	
	coeff	se	coeff	se	coeff	se
fs	-0.0132	0.0421	0.1341	0.0444	0.0640	0.0302
ms	0.0993	0.0493	-0.0541	0.0496	0.0345	0.0344

TABLE 5. Direct effects of parents' schooling by proxying U.

5. Concluding Remarks

Within the broad issue of education transmission there are two important recent lines of research related to the present work. One analyzes the *total causal effect* of parents' education on children's controlling for parents' unobserved endowment, by use of twin parents, adoptees or compulsory schooling law instruments ([1, 7, 8, 9, 25, 31]). The other, lead by Heckman ([16, 17, 23]) studies the effect of sequential interventions and their complementarity on the *evolution* of endowments and on labor market outcomes.

By applying recently developed finite mixture models to the NCDS dataset, the present paper controls for the child's own schooling endowments at the age of 16, and measures the *direct causal effect* of parents' schooling on children's educational achievement at the same age. The effect of parents' schooling on children's educational attainment *given* the latter's potential mainly reflects parental pressure, and can thus be interpreted as a 'role' effect. To allow for the possibility of its dependence on gender we consider sons and daughters subsamples separately. We find that only fathers' education matters, and that its impact is entirely confined to the education of their sons.

This result may reflect social norms of Western families in the nineteen-seventies (the data domain), and if the women's role has changed a different picture may emerge from more recent data. But the message we get from our findings remains that children respond to family pressure on schooling attainment. From a policy – or rather 'cultural'– viewpoint this suggests that when parents' pressure is weak only the social environment, school primarily, can make up for this loss by helping the young to appreciate the value of education.

6. Appendix

6.1. **Direct and indirect causal effect.** Consider the following causal diagram where Y is a binary outcome, X is a discrete input of interest, U and V are discrete unobservable variables and Z is a covariate:

$$\begin{tikzcd}
 & V \arrow[r] \arrow[dl] \arrow[d] & X \arrow[dl] \arrow[d] \\
Z \arrow[r] & U \arrow[r] & Y
\end{tikzcd}$$

The corresponding model may be seen as a simplified version of the one stated in (1) and (2) where Y represents scholastic attainment and X parent's years of schooling while U and V are meant to capture respectively child's and parent's unobservable endowments and Z represents a family background variable affecting U and which may also depend on V.

In the following, for conciseness and when no ambiguity arises, we use the convention that a conditioning variable is denoted by its value, thus for example $\Pr(Y \mid x, v)$ means $\Pr(Y \mid X = x, V = v)$. If we could control for Z and V, the causal effect of X on Y could be measured by

$$\Delta_x(z, v) = \Pr(Y = 1|x + 1, z, v) - \Pr(Y = 1|x, z, v),$$

This may be expanded by using the conditional independence of Y and V given X, U which implies that

$$\Pr(Y|x, z, v) = \sum\nolimits_u \Pr(Y \mid x, u) \Pr(U = u \mid x, z, v);$$

now let $P_{x,u} = \Pr(Y = 1|x, u)$ and $Q_{u|x,z,v} = \Pr(U = u|x, z, v)$ and write

$$\Delta_x(z, v) = \sum\nolimits_u (P_{x+1,u}Q_{u|x+1,z,v} - P_{x,u}Q_{u|x,z,v}).$$

By adding and subtracting $\sum_u P_{x,u}Q_{u|x+1,z,v}$, rearranging terms and noting that if U has m levels we may write $Q_{1|x,z,v} = 1 - \sum_2^m Q_{u|x,z,v}$, one gets

$$\Delta_x(z,v) = \sum_{u=1}^{m}(P_{x+1,u} - P_{x,u})Q_{u|x+1,z,v} + \sum_{u=2}^{m}(P_{x,u} - P_{x,1})(Q_{u|x+1,z,v} - Q_{u|x,z,v}).$$

The first component is an average of the direct effect of X on Y weighted with the conditional distribution of U; in the paper we report the direct effect conditional on each value of U. The second term is the sum of the products of the effect on Y when U changes from a reference category to a given category, times the effect of X on U and may be interpreted as a measure of the indirect effect of X on Y carried to Y through U. Though the above decomposition is not unique since, for instance, in the first component $Q_{u|x,z,v}$ could be used as weights with minor changes in the second component, it is one way of generalizing the decomposition which holds in the system of linear equations

$$\begin{aligned} Y &= a_0 + a_1X + a_2U + \epsilon, \\ U &= b_0 + b_1X + b_2Z + b_3V + \eta. \end{aligned}$$

where the total effect is $a_1 + b_1a_2$ and the direct effect is a_1.

The causal diagram above implies also that, if we do not control for V, the effect of Z on U may be biased. This can be made explicit by an expansion similar to the one given above where we assume that V has r distinct levels

$$\begin{aligned} &\Pr(U \mid z+1) - \Pr(U \mid z) = \sum_v [\Pr(U \mid z+1, v) - \Pr(U \mid z, v)] \Pr(V = v \mid z+1) \\ &+ \sum_2^r [\Pr(U \mid z, v) - \Pr(U \mid z, 1)][\Pr(V = v \mid z+1) - \Pr(V = v \mid z)] \end{aligned}$$

where the first component is simply an average causal effect of Z on U weighted with the distribution of V while the second component is 0 whenever Z and V are independent, otherwise it adds a bias term whose sign depends on the direction of the dependence between Z and V.

6.2. **EM algorithm and information matrix.**

6.2.1. *Likelihood inference.* **The true log-likelihood.** Let $\boldsymbol{n}(i)$ be the 2^7 vector containing the frequency table of the response variables $\boldsymbol{Y}$ in lexicographic order for the subjects with covariate $\boldsymbol{x}_i$; if there is a single subject with such features, $\boldsymbol{n}(i)$ is a vector of 0s except for a 1 in the cell corresponding to the response pattern $\boldsymbol{y}(i)$. Let also $\boldsymbol{q}(i)$ denote the probability distribution for subjects with covariate $\boldsymbol{x}_i$. The log-likelihood may be written as

$$L = \sum L_i = \sum \boldsymbol{n}(i)' \ln[\boldsymbol{q}(i)].$$

Maximizing this log-likelihood is as a problem of incomplete data which may be tackled by the EM algorithm (Dempster *et al.* [19]).

The latent log-likelihood. Let $\boldsymbol{L} = (\boldsymbol{1}_{m+1}{}' \otimes \boldsymbol{I}_{2^7})$ denote the matrix which marginalizes with respect to the latent variable U, $\boldsymbol{p}(i)$ the vector containing the joint probability distribution of $(U, \boldsymbol{Y})$ for subjects with covariate $\boldsymbol{x}_i$ and $\boldsymbol{m}(i)$ the vector containing the unobservable frequency table of $(U, \boldsymbol{Y})$ for subjects with covariate $\boldsymbol{x}_i$. Clearly $\boldsymbol{n}(i) = \boldsymbol{L}\boldsymbol{m}(i)$ and $\boldsymbol{q}(i) = \boldsymbol{L}\boldsymbol{p}(i)$. If the latent class U could be observed, the corresponding log-likelihood would have the form

$$\Lambda = \sum \Lambda_i = \sum \boldsymbol{m}(i)' \ln[\boldsymbol{p}(i)].$$

The E step. Because the multinomial is a member of the exponential family, the conditional expectation involved in the E step is equivalent to computing the so called posterior probability of latent class U given the observed configuration $\boldsymbol{y}$

$$\Pr(U = u \mid \boldsymbol{y}_i, \boldsymbol{x}_i) = \frac{p_{u,\boldsymbol{y}}(i)}{q_{\boldsymbol{y}}(i)}$$

so that $m_{u,\boldsymbol{y}}(i) = n_{\boldsymbol{y}}(i)\Pr(U = u \mid \boldsymbol{y}_i, \boldsymbol{x}_i)$ follows from a simple expectation of a multinomial distribution for U.

The M step. Implementation of the method of scoring for the maximization of Λ with respect to the model parameters $\boldsymbol{\psi}$ requires computation of the score vector (first derivative with respect to $\boldsymbol{\psi}$) and of the expected information matrix (minus the expected value of the second derivative). Since Λ is a multinomial log-likelihood, exponential family results can be exploited to make such calculations straightforward. In practice, after rewriting Λ in terms of the canonical parameters of the multinomial distribution, say $\boldsymbol{\theta}(i)$, there are invertible and differentiable mappings from $\boldsymbol{\theta}(i)$ to the vector of probabilities $\boldsymbol{p}(i)$ and from $\boldsymbol{p}(i)$ to $\boldsymbol{\lambda}(i)$ (the latter mapping is described after Proposition 1), while $\boldsymbol{\lambda}(i)$ is linked to $\boldsymbol{\psi}$ by the linear regression model (7). The interested reader may see Dardanoni–Forcina [18] or Bartolucci *et al.* [3] for details.

6.2.2. *Computational issues.* The EM algorithm is a very robust method of estimation of the model parameters for latent class models. However, it suffers from at least two drawbacks: it can be very slow with large data sets, and, by itself, does not provide a consistent estimate of the variance-covariance matrix of the model parameters. This is so because the expected information matrix of the latent likelihood is based on the assumption that the vector $\boldsymbol{m}(i)$ is known, using its inverse as an estimate of the variance matrix implies that standard errors will in general be underestimated. The correct information matrix may be computed by differentiating the incomplete data likelihood as follows. Write $L_i = \boldsymbol{n}(i)'\tilde{\boldsymbol{G}}\boldsymbol{\gamma}_i - n_i \ln[\boldsymbol{1}' \exp(\tilde{\boldsymbol{G}}\boldsymbol{\gamma}_i)]$ where $\boldsymbol{\gamma}_i$, the canonical parameter of the observed multinomial, may be written as $\tilde{\boldsymbol{H}} \ln[\boldsymbol{L} \exp(\boldsymbol{G}\boldsymbol{\theta}_i)/\boldsymbol{1}' \exp(\boldsymbol{G}\boldsymbol{\theta}_i)]$, where $\tilde{\boldsymbol{H}}$ is a $t \times (t-1)$ contrast matrix used to define the canonical parameters and $\tilde{\boldsymbol{G}}$ is its right inverse while $\boldsymbol{G}$ is the design matrix which defines the canonical parameters $\boldsymbol{\theta}$ for the latent distribution $\boldsymbol{p}(i)$ which has v columns of full rank. By differentiating

L_i by the chain rule with respect to $\boldsymbol{\psi}$ one may write

$$\sum \frac{\partial L_i}{\partial \boldsymbol{\psi}} = \boldsymbol{B}_i{}'\boldsymbol{R}_i{}'\boldsymbol{G}'\boldsymbol{\Omega}_i\boldsymbol{L}'diag(\boldsymbol{q}_i)^{-1}\tilde{\boldsymbol{H}}'\tilde{\boldsymbol{G}}'(\boldsymbol{n}(i) - n_i\boldsymbol{q}_i)$$

where $n_i = \boldsymbol{1}'\boldsymbol{n}(i)$, $\boldsymbol{\Omega}_i = diag[\boldsymbol{p}(i) - \boldsymbol{p}(i)\boldsymbol{p}(i)']$ and $\boldsymbol{R}_i$ is the derivative of the canonical parameter $\boldsymbol{\theta}_i$ with respect to $\boldsymbol{\lambda}_i{}'$. Because $E(\boldsymbol{n}(i) - n_i\boldsymbol{q}_i) = \boldsymbol{0}$, minus the expected value of the second derivative may be written as

$$\boldsymbol{F}_i = \boldsymbol{B}_i{}'\boldsymbol{R}_i{}'\boldsymbol{G}'\boldsymbol{\Omega}_i\boldsymbol{L}'diag(\boldsymbol{q}_i)^{-1}\tilde{\boldsymbol{H}}'\tilde{\boldsymbol{G}}'\boldsymbol{L}\boldsymbol{\Omega}_i\boldsymbol{G}\boldsymbol{R}_i\boldsymbol{B}_i$$

(where $\tilde{\boldsymbol{H}}'\tilde{\boldsymbol{G}}'$ is simply equal to $\boldsymbol{I}_t - \boldsymbol{1}_t\boldsymbol{1}_t{}'/t$ with $t = 2^7$), so the information matrix is simply $\sum_i \boldsymbol{F}_i$.

6.3. **Model identifiability.** Formally identifiability concerns the mapping from the manifest probability distribution $\boldsymbol{q}$ and the model parameter $\boldsymbol{\psi}$ (Rothemberg, [32]). Recall that a model is globally identified when there are no two points in the parameters' space, say $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, such that $\boldsymbol{q}(\boldsymbol{\psi}_1) = \boldsymbol{q}(\boldsymbol{\psi}_2)$. However, the weaker notion of local identifiability (see Catchpole and Morgan, [13]) is of more direct statistical interest and also easier to verify. It requires that at any $\boldsymbol{\psi}_0$ the set of points such that $\|\boldsymbol{q}(\boldsymbol{\psi}) - \boldsymbol{q}(\boldsymbol{\psi}_0)\| = 0$ satisfy $\|\boldsymbol{\psi} - \boldsymbol{\psi}_0\| > \delta > 0$. This implies that there exist no parameter value with a neighborhood where the likelihood is constant and thus the information matrix must be positive definite everywhere.

To analyze the identifiability of our model, let the vector $\boldsymbol{\gamma}$ obtained by stacking the vectors $\boldsymbol{\gamma}_i$ (the vectors of canonical parameters of the saturated log-linear model for each subject in the manifest distribution) and consider the different parametric transformations involved:

(1) from $\boldsymbol{\gamma}$ to $\boldsymbol{\theta}$, the vector obtained by stacking the vectors of canonical parameters of the latent class model for each subject,

(2) from $\boldsymbol{\theta}$ to $\boldsymbol{\lambda}$, the vector obtained by stacking the vectors of marginal parameters for each subject,

(3) the regression model $\boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{\psi}$.

Identifiability of the regression model is easily established by checking that $\boldsymbol{B}$ is of full column rank. Results from Bartolucci *et al.* [3] ensure that the transformation to the marginal parameters is invertible and differentiable. So, the crucial transformation is the first one. However, as shown by Forcina [21], the regression model can make identifiable a model which is not, simply because estimating the full $\boldsymbol{\lambda}$ vector is much more demanding than estimating $\boldsymbol{\psi}$. Though, due to the complexity of our model, no analytic result is available to check local identifiability, full rank of the jacobian of the mapping from $\boldsymbol{q}$ to $\boldsymbol{\psi}$ may be tested numerically in a fast and efficient way as described by Forcina [21] for a wide range of values sampled at random. Since in our case in 10000 runs no instance was detected where the rank of the jacobian was any close to being deficient, we may conclude that our model is identifiable for a wide range of parameter values.

	Daughters		Sons	
	coeff	se	coeff	se
$\boldsymbol{\beta}_{OL}$				
father sch	-0.0173	0.0454	0.1150	0.0488
mother sch	0.0831	0.0532	-0.0363	0.0549
$\boldsymbol{\beta}_U(U=2)$				
father sch	0.1612	0.0460	0.2612	0.0478
mother sch	0.2052	0.0546	0.1164	0.0566
father int	0.6236	0.1615	0.5018	0.1669
mother int	0.2339	0.1576	0.4507	0.1612
$\boldsymbol{\beta}_U(U=3)$				
father sch	0.2242	0.0701	0.1371	0.0674
mother sch	0.0969	0.0736	0.2220	0.0707
father int	1.2581	0.1860	1.2041	0.1653
mother int	1.1202	0.1880	0.5648	0.1661
Other parameters				
$\alpha_{OL}(U=3)$	2.4018	0.2233	2.2796	0.2372
$\alpha_{OL}(U=2)$	0.5587	0.1269	0.2160	0.1166
$\alpha_{OL}(U=1)$	-1.8738	0.1477	-2.5745	0.1919
$\alpha_{LM}(U=3)$	3.2083	0.5573	5.1050	1.3935
$\alpha_{LM}(U=2)$	-0.1648	0.1364	0.9776	0.1527
$\alpha_{LM}(U=1)$	-2.7204	0.21184	-2.1853	0.1797
γ_{LM}	0.4330	0.1933	0.1243	0.1938
$\alpha_{LR}(U=3)$	1.9836	0.2930	2.8633	0.5183
$\alpha_{LR}(U=2)$	-0.2676	0.1390	0.0489	0.1148
$\alpha_{LR}(U=1)$	-3.2065	0.2907	-2.1427	0.1406
γ_{LR}	0.8251	0.1573	1.0570	0.1661
$\alpha_{EM}(U=3)$	1.5871	0.1505	2.3560	0.2495
$\alpha_{EM}(U=2)$	-0.7403	0.1386	-0.4214	0.1172
$\alpha_{EM}(U=1)$	-3.3926	0.2864	-3.1687	0.2467
$\alpha_{ER}(U=3)$	1.6289	0.1398	1.6389	0.1551
$\alpha_{ER}(U=2)$	-0.2881	0.1236	-0.6922	0.1203
$\alpha_{ER}(U=1)$	-3.5126	0.3340	-3.3828	0.2617
$\alpha_{NCP}(U=3)$	0.1884	0.0799	-0.3685	0.0823
$\alpha_{NCP}(U=2)$	-0.4056	0.0905	-0.8411	0.0888
$\alpha_{NCP}(U=1)$	-1.0264	0.0866	-2.0463	0.1200
$\alpha_{NCS}(U=3)$	0.1300	0.0792	-0.4009	0.0826
$\alpha_{NCS}(U=2)$	-0.4268	0.0934	-0.8843	0.0890
$\alpha_{NCS}(U=1)$	-1.7460	0.1161	-1.8259	0.1077
$\alpha_U(U=2)$	0.5145	0.1270	0.7054	0.1253
$\alpha_U(U=3)$	-0.3973	0.1294	-0.1756	0.1056

TABLE 6. Parameters' estimates for system (6).

References

[1] Antonovics, K. L. and A. S. Goldberger (2005): "Does Increasing Women's Schooling Raise the Schooling of the Next Generation? Comment", *American Economic Review* **95**, pp. 1738–1744

[2] Arcidiacono, P. and J. B. Jones (2003): "Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm", *Econometrica* **71**, pp. 933–946

[3] Bartolucci, F., R. Colombi and A. Forcina (2006): "An Extended Class of Marginal Link Functions for Modelling Contingency Tables by Equality and Inequality Constraints", *Statistica Sinica*, to appear

[4] Bartolucci, F. and A. Forcina (2005): "Likelihood Inference on the Underlying Structure of IRT Models", *Psychometrika* **30**, pp. 140–159

[5] Bartolucci, F. and A. Forcina (2006): "A cCass of Latent Marginal Models for Capture-Recapture Data with Continuous Covariates", *Journal of the American Statistical Association* **101**, pp. 786–794

[6] Bergsma, W. and T. Rudas, (2002): "Marginal Models for Categorical Data" *Annals of Statistics* **30**, pp. 140–159

[7] Behrman, J. R. and M. R. Rosenzweig (2002): "Does Increasing Women's Schooling Raise the Schooling of the Next Generation?", *American Economic Review* **92**, pp. 323–334

[8] Behrman, J. R. and M. R. Rosenzweig (2005): "Does Increasing Women's Schooling Raise the Schooling of the Next Generation? Reply", *American Economic Review* **95**, pp. 1745–1751

[9] Black, S. E., P. J. Devereux and K. G. Salvanes (2005): "Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital", *American Economic Review* **95**, pp. 437–449

[10] Blundell, R., L. Dearden and B. Sianesi (2005): "Evaluating the Effect of Education on Earnings: Models, Methods and Results from the National Child Development Survey", *Journal of the Royal Statistical Society A*, **168**, pp. 473–512

[11] Boudon, R. (1974): *Education, Opportunity, and Social Inequality: Changing Prospects in Western Society* Wiley, New York

[12] Cameron, S. V., and J. J. Heckman (1998): "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males", *Journal of Political Economy* **106**, pp. 262–333

[13] Catchpole, E. A., and Morgan, B. J. T. (1997): "Detecting Parameter Redundancy," *Biometrika*, 84, pp. 187-196.

[14] Chalak, K., and H. White (2007): "An Extended Class of Instrumental Variables for the Estimation of Causal Effects ", *submited*

[15] Colombi, R. and A. Forcina (2001): "Marginal Regression Models for the Analysis of Positive Association", *Biometrika* **88**, pp. 1007–1019.

[16] Cunha, F. and J. J. Heckman (2007): "The Technology of Skill Formation", *American Economic Review* **97**, pp. 31-47

[17] Cunha, F., J. J. Heckman and S. Schennach (2007): "Estimating the Technology of Cognitive and Noncognitive Skill Formation", Working Paper

[18] Dardanoni, V. and A. Forcina (2008): "Multivariate ordered regression", *mimeo*

[19] Dempster, A. P., N. M. Laird and D. B. Rubin (1977): "Maximum Likelihood for Incomplete Data Via the EM Algorithm", *Journal of the Royal Statistical Society* Series B **39**, pp. 1–22

[20] Erikson, R., J. H. Goldthorpe, M. Jackson, M. Yaish, and D. R. Cox (2005): "On Class Differentials in Educational Attainment", *Proceedings of the National Academy of Science* **102**, pp. 9730–9733

[21] Forcina, A. (2008): "Identifiability of Extended Latent Class Models with Individual Covariates," *Computational Statistics and Data Analysis*, 52, pp. 5263-5268

[22] Goodman, L. (1974): "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models", *Biometrika* **61**, pp. 215–231

[23] Heckman, J., J. Stixrud and S. Urzùa (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior", *Journal of Labor Economics* **24**, pp. 411-482

[24] Heckman, J. and B. Singer (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, **52**, pp. 271-320

[25] Holmlund, H., Lindahl, M. and E. Plug (2008): "The Causal Effect of Parent's Schooling on Children's Schooling: A Comparison of Estimation Methods, *IZA Discussion Paper*, **3630**.

[26] Huang G. and K. Bandeen-Roche (2004): "Building an Identifiable Latent Class Model, with Covariate Effects on Underlying and Measured Variables" *Psychometrika* **69**, pp. 5-32

[27] Keane, M. P. and K. I. Wolpin (1997): "The Career Decisions of Young Men", *Journal of Political Economy* **105**, pp. 473-522

[28] Keane, M. P. and K. I. Wolpin (2001): "The Effect of Parental Transfers and Borrowing Constraints on Educational Attainment", *International Economic Review* **42**, pp. 1051-1103

[29] Lindsay, B., C. Clogg, and J. Grego (1991): "Semiparametric Estimation of the Rasch Model and Related Exponential Response Models, Including a Simple Latent Class Model for Item Analysis", *Journal of the American Statistical Association* **86**, pp. 96-107

[30] Pearl, J. (2000): *Causality*, Cambridge University Press

[31] Plug, E. (2004): "Estimating the Effect of Mother's Schooling Using a Sample of Adoptees", *The American Economic Review* **94**, pp. 358-368

[32] Rothenberg, T. J. (1971): "Identification in Parametric Models," *Econometrica*, **39**, pp. 577-591.

[33] Vermunt, J. K. and J. Magidson (2003), "Latent Class Analysis", in *Encyclopedia of Research Methods for the Social Sciences*, Michael S. Lewis-Beck, Alan Bryman and Tim Futing Liao editors, Sage Publications, NewBury Park.

[34] Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, The MIT Press.

Facoltà di Economia, Università di Palermo
E-mail address: `vdardano@unipa.it`

Facoltà di Economia, Università di Perugia
E-mail address: `forcina@stat.unipg.it`

Facoltà di Economia, Università di Palermo
E-mail address: `modica@unipa.it`