YOUNG LIVES STUDENT PAPER

The Dynamics of Skill Formation in a Developing Country: Evidence from Ethiopia

Ibirénoyé Honoré Romaric Sodjahin

May 2017

Thesis submitted in part fulfilment of the requirements for the degree of MPhil in Economics for Development, Paris University, Sorbonne School of Economics.

The data used come from Young Lives, a longitudinal study of childhood poverty that is tracking the lives of 12,000 children in Ethiopia, India (in the states of Andhra Pradesh and Telangana), Peru and Vietnam over a 15-year period. **www.younglives.org.uk**

Young Lives is core-funded by UK aid from the Department for International Development (DFID).

The dynamics of skill formation in a developing country : evidence from Ethiopia

Presented and Defended by: Ibirénoyé Honoré Romaric Sodjahin*

Université Paris 1 – *École d'Économie de la Sorbonne*

Master 2 *Recherche Development Economics*

Supervised by: Pr. Tanguy Bernard†

May 19^{th}, 2017

Abstract

There is a consensus in the community of researchers that ability explains a substantial part of the differences across people of success in socioeconomic life, and that ability gaps across people emerge at childhood before they start school. Building on recent advances in the child development literature in Economics pioneered by James J. Heckman, we estimate two transitions technology of skill formation using four waves of survey data in Ethiopia which are part of the longitudinal project "Young lives" funded by the UK aid department.

We found evidence that early life conditions, including antenatal care have significant effects on child' health, and that child health is positively related to a higher

*Many thanks to Pr. Tanguy Bernard for supervision, constant support and feedback. Also grateful to Pr. David Margolis and Pr. Marc-Arthur Diaye for advices and very useful hints, and to the UK data archive who granted us access to Young Lives project data. All errors are mine.

†University of Bordeaux IV and International Food Policy Research Institute (IFPRI)

level of ability in the four rounds, except in the last round where we observed a negative effect of child health on noncognitive skills only. We also found evidence of self-productivity for cognitive skills and noncognitive skills and cross productivity from cognitive to noncognitive skills.

We don't find any effect of parental investment on child's cognitive/noncognitive skills and even a negative effect at age 5. This is partly due to a huge missing data which leads us to exclude some important variables from the analysis.

Keywords: Children, cognitive skills, non-cognitive skills, state space models

JEL: J13, O15, C32

L'université de Paris 1 Panthéon-Sorbonne n'entend donner aucune approbation ni désapprobation aux opinions émises dans ce mémoire: elles doivent être considérées comme propre à leur auteur.

The University of Paris 1 Panthéon-Sorbonne neither approves nor disapproves of the opinions expressed in this dissertation: they should be considered as the author's own.

Contents

1 Introduction **1**

2 Empirical strategy **7**

2.1 Latent variable estimation procedure 7

3 Data **11**

3.1 Introduction 11

3.2 Observed variables for latent factor estimation 12

3.3 Missing data issues 14

4 Results **16**

5 Conclusion **21**

6 Appendix **i**

1 Introduction

Why do some individuals succeed better in their socio-economic life while others succeed less ? This challenging question received much attention from economists quite recently, building in the advance in child development literature. Indeed, since the end of the world war II, economists were rather interested in models explaining differences across countries in per capita income[1] at the aggregate level. In a survey of the state of the art, Cunha & Heckman (2007) claimed this : *It is now well documented*[2] *that people have diverse abilities, that these abilities*[3] *account for a substantial portion of the variation across people in socioeconomic success, and that persistent and substantial ability gaps across children from different socioeconomic groups emerge before they start school*. In addition, Cunha et al. (2010, p. 884) stress the importance of cognitive skills in producing socioeconomic success, and the fact that noncognitive skills (that is personality, social and emotional traits) were as important as cognitive skills to explain differences across people in outcomes (Koch et al. 2015). With that in mind, a way to understand the differences of socioeconomic outcomes across people is to identify how cognitive and noncognitive skills are formed over time, starting from early childhood; then one can provide policy recommendations on later remediation targeting the disadvantaged children.

The theoretical framework used in skill formation analysis so far considered childhood as a single stage[4]; this implies that inputs used to produce skills are perfect substitutes. In order to take into account most of the recurrent results obtained in empirical research, Cunha & Heckman (2007) proposed a model of skill formation with multiple stages of childhood, where inputs at different stages are complements. Based on their

[1] Along these lines, we can quote early works as the Harrod(1948) and Domar(1947)'s models, the Solow(1956) and Swan(1956)'s models, and the Ramsey(1928) model, extended by Cass(1965) and Koopman(1965). The per capita income is important because it is a rough picture of the standard of living of a country. Later works in the same line are the Romer(1986) model, the Lucas(1988) model, the Aghion and Howitt(1992) model.

[2] See the references they quoted.

[3] In the child development literature, the words "ability" and "skill" refer to the same thing, and thus will be used interchangeably.

[4] See the references cited in Cunha & Heckman (2007)

theoretical framework, this thesis aims first to test the predictions of self-productivity and cross-productivity, which, put together, explain why skill begets skill through a multiplier process :

- Skills (cognitive/noncognitive) are self-productive : this prediction refers to the idea that the skills produced at one stage augment the skills attained at later stages; that is skills are self-reinforcing over time at different stages;

- Skills (cognitive/noncognitive) are cross-productive : this prediction relies on the idea that cognitive skills at time t raises noncognitive skills at time $t+1$ and noncognitive skills at time t increases cognitive skills at time $t+1$.

Cunha & Heckman (2007) also introduced the notions of critical periods and sentitive periods in the formation of skills :

- A stage t^* is said to be critical for a skill if this stage is **the only one** in which self-productivity can occur during childhood;

- A stage t^+ is said to be sensitive for a skill if the increase of skills from a period t to $t+1$ is **the highest** when $t = t^+$

Determining the sensitive and the critical periods can help policymakers to act appropriately and efficiently when implementing policies targeting disadvantaged children. This thesis will also investigate the effect of other factors like the child's health in the process of child's skills accumulation.

Previous works found evidence of self-productivity and cross-productivity of skills. It is the case of Cunha et al. (2010) who used the National Longitudinal Survey of Youth (NLSY79), estimated the technology of skill formation on a sample of 2207 first born white children in 1986; Cunha & Heckman (2008) and Helmers & Patnam (2011) also reach similar conclusions. Todd & Wolpin (2007) use the same dataset like Cunha et al. (2010) and focused on the sources of test score gaps between black, white, and hispanic

children; they found the widening of minority-white test score gaps with age and differences in the gap pattern between hispanics and blacks. Thuilliez et al. (2010) found that malaria was negatively correlated with cognitive performance in Mali. More recently, Bono et al. (2016) used longitudinal survey data from the UK Millennium Cohort Study and found that maternal time investment is a quantitatively important determinant of skill formation.

As pointed out by Helmers & Patnam (2011), so far the findings about the development process of a child's abilities have been obtained using survey data from rich countries, but little is known for the implications of these findings in developing countries' context, because preferences are not the same in the two contexts, and in addition, credit constraints is an issue in developing countries. There are two main influences shaping a child's abilities during his multistage development process : his genetic endowment and inputs received from his environment. It is acceptable to think that genetic endowments might express themselves depending on the family environment, parental care and investment[5]. So, empirical results obtained in high income countries might diverge from those in low income countries. The only study, from the best of my knowledge, investigating the dynamic of skill formation for a developing country is Helmers & Patnam (2011) who estimate a single transition technology for India. My thesis aims to estimate a two transition which is more informative and made possible thanks to the availability of more data, and allows an analysis from early childhood until adolescence. Another contribution of my thesis is the fact that unlike Cunha et al. (2010), we have prior information of children during their early childhood, including the antenatal conditions, while the sample used by Cunha et al. (2010) has information on children from 6 years old onwards. My dataset thus allows to study the effect of early parental investment in the dynamics of skill formation. Moreover, two cohorts of children are surveyed : a

[5]As noted by Hunter (2008), the behaviour of our genes can be altered by experience; this much we can tell by observing identical twins, who over time tend to diverge both physiologically (developing differences in, say, height and posture) and psychologically (exhibiting different personality traits and even, sometimes, sexual orientations).

young cohort and a old cohort. In the fourth wave of survey conducted in 2014, the old cohort was aged around 22. The older cohort could be used to test the robustness of results obtained on young cohorts.

We use data from the Ethiopia part of the Young Lives project, a long-term study of childhood poverty being carried out in 4 countries : India, Peru, Vietnam and of course Ethiopia. The broad objective of the *Young Lives* project is to improve understanding of the causes and consequences of childhood poverty and to examine how policies affect children's well-being. Extensive child, household and community level questionnaires are administered to capture information on various aspects of the child's life including household demographics, caregiver background, child health (both physical and mental), economic shocks, household consumption, as well as social, economic and environmental context of each community. In Round 1, 2000 children aged around one (the "younger" cohort) and 1000 children aged around eight (the "older" cohort) were surveyed in 2002. Following up, Round 2 involved tracking the same children and surveying them in 2006 at age five and twelve respectively. The subsequent follow-up surveys occur in 2009 and 2014.

One of the big challenges is to find an appropriate measure of what *cognitive skills*, *non-cognitive skills*, *parental investment* or health are. As noted by Blume et al. (2010), economic theory does not dictate the appropriate empirical measures of contextual variables that a researcher ought to use. Given that cognitive/non-cognitive skills, parental input and health are in principle unobserved, they will be treated as latent variables for which observed indicators should be associated with. The measures used to capture cognitive skills aim to capture a child's general intelligence and his ability to solve abstract problems. As such, measures for cognitive skills differ from those for non-cognitive skills which represent aspects of a child's personality, including timidity, extraversion, motivation, self-confidence. Following previous works, we used test scores[6] to estimate

[6]writing, reading, Raven Progressive Matrices, numeracy and Peabody Picture Vocabulary, Cognitive Development Assessment Quantitative Test, Early Grade Reading Assessment

factor scores for cognitive ability at different ages.

For non-cognitive skills, previous works made use of these measurement indicators:

- the Child Mental Ability Indicators from the Strengths and Difficulties Questionnaire[7];
- the child Personality Measures indicated from questions rated on the Likehart Scale by child[8];
- the level of fluency and communication in native language;
- performance in pre-school (interactive and social nature);
- does child travel to school with friends, parents or alone ?

We follow the procedure[9] in (Helmers & Patnam 2011) and write child's skill level at age *t* as a function of the child's past level of skills, current parental investment, and other contemporaneous variables including child, caregiver, and household characteristics:

$$\theta_t^\kappa = f(\theta_{t-1}^\kappa, \theta_t^I, X_t)$$

where :

1. θ_t^κ denotes a child's skill level of skill κ for age *t* (where $\kappa \in \{Cognitive, Non-cognitive\}$ hereafter C and N);
2. θ_t^I denotes parental investment;
3. X_t denotes a vector of child, caregiver and household characteristics.

First of all, skills (cognitive and noncognitive) and parental investment will be estimated using observed proxies thanks to a dynamic factor model (a state space model)

[7]Emotional conduct, does child work at home or outside, pro-social behaviour, conduct problems
[8]Friendliness, pride, determination, social trust, group membership
[9]This procedure is itself inspired from (Cunha & Heckman 2007,8, Cunha et al. 2010).

to produce latent variable for each of them. These latent variables will be used in the subsequent analysis.

The evidence of self-productivity and cross-productivity of abilities will be shown by the significance and the positive sign of the coefficient associated to θ_{t-1}^{κ} in the previous equation. In particular, self-productivity is the link between cognitive skills at time $t-1$ and cognitive skills at time t, while cross-productivity is the link between cognitive skills at time $t-1$ and noncognitive skills at time t and vice versa. If properly measured, we expect parental investment at time $t-1$ to be positively related to cognitive/noncognitive skills at time t.

The intuition raised by the model developed in Cunha & Heckman (2007) is that cognitive and noncognitive skills accumulation are interdependent processes and should be modelled jointly. In addition, as pointed out by Helmers & Patnam (2011), the heterogeneity of the quality of parental investment (which is unobserved) should differently influence the child's skills. An illustration of this point is made by Jane Waldfogel[10] in her book whose title is : *What do children need ?*. She argues that "maternal sensitivity is the most important predictor of child social and emotional development". Bono et al. (2016) showed using large longitudinal survey data from the UK Millennium Cohort Study that maternal time is a quantitatively important determinant of skill formation and that its effect declines with child age.

Because parental investment is likely to be endogenous, we will use an instrumental variable strategy to account for this endogeneity. Helmers & Patnam (2011) used for India specific shocks which affect household's wealth and a child's birth order as instruments. Validity of these instruments rests on the assumption that their effect on child outcomes works exclusively through parental investment conditional on a set of control variables. This appears to be a credible assumption as children will be affected by unexpected shocks to household wealth only through adjustments made by parents in their investment in their children. We expect that the instrumental variable strategy and the

[10] cited by Helmers & Patnam (2011, p. 254)

number of control variables used will be enough to account for the changing nature of children initial endowments.

2 Empirical strategy

2.1 Latent variable estimation procedure

This section uses the formalism presented in a rather recent book by Durbin & Koopman (2012)[11].

The estimation of the latent variables relies on a linear state space model. The main purpose of state space analysis is to infer the relevant properties of the state (which is unobserved latent variable) given some observed variables. The general linear gaussian state space model can be written in the form :

$$\begin{aligned} \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t & \eta_t &\sim N(0, Q_t) \\ y_t &= Z_t\alpha_t + \epsilon_t & \epsilon_t &\sim N(0, H_t),\ with\ t = 1, \cdots T \end{aligned} \tag{1}$$

α_t is an unobserved $m \times 1$ vector called the *state vector*. It implies that we have m state equations.

y_t is a $p \times 1$ vector of observations called the *observation vector*. It implies that we have p observation equations.

The matrices T_t, R_t, Q_t, Z_t and H_t are initially assumed to be known according to Durbin & Koopman (2012), but Gourieroux & Monfort (1995) use the term *Nonrandom*. The two ideas are equivalent.

The error terms η_t and ϵ_t are assumed to be serially independent and independent of each other at all time points. The initial state vector α_1 is assumed to be $N(a_1, P_1)$ independently of $\epsilon_1, \cdots, \epsilon_T$ and $\eta_1, \cdots, \eta_T$.

In our estimation, we consider models with one state equation ($m = 1$) because we are only interested in one latent factor. The number of measurement equations depends on the number of observed proxies variables for each latent factor. For example, the

[11] Another interesting presentation can be found in Gourieroux & Monfort (1995, chap. 13).

latent factor model for health uses two measurement equations associated to the two variables : the height-for-age z-score and the weight-for-age z-score which are growth standards of the World Health Organization.

The next step consists of evaluating the likelihood function of the model, that is the density of the observations $y_1, y_2, \cdots, y_T$ viewed as a function of the unknown parameters of the model (note that the matrices T_t, R_t, Q_t, Z_t and H_t have to be estimated), keeping in mind that y_1 is a vector of dimension p, where each value is the first observation of each observed proxy variable.

If we denote the set of initial parameters[12] by $\theta^{(0)}$, then an iteration formula will allow to compute the next vector of parameters $\theta^{(1)}$. The iteration formula depends of the numerical optimization procedure used (eg. Newton-Raphson) and a stopping rule has to be set for the algorithm to terminate; it might be the maximum number of iterations or a tolerance (the difference between two consecutive values of the likelihood).

It is advocated to use the Newton-Raphson algorithm when the likelihood function is concave. Sometimes, the Newton-Raphson algorithm is very slow to converge[13]. Then I use a mixture of algorithms, in particular I start with the Broyden–Fletcher–Goldfarb–Shanno (BFGS)[14] and end up with the Newton-Raphson algorithm.

For the choice of the initial vector of parameters $\theta^{(0)}$, the fact that I have no prior knowledge on these parameters leads me to use the diffuse Kalman filter, a procedure for searching the initial values proposed by De Jong (1991).

Although this could be restrictive, I assume for simplicity that the errors terms in the equations of the cognitive and the noncognitive skills are uncorrelated, and that the only source of correlation is due to observable characteristics. With that said, the two equations can be estimated separately by an instrumental variables procedure. Indeed, the parental investment is likely to be endogenous. Parent might invest on their children

[12] which are the elements of the nonrandom matrices referred to above plus the initial value of the state variable and its variance-covariance matrix

[13] I experienced more than 100 000 iterations without convergence

[14] which seems to speed up the convergence process

because they have some expectations that the children will take care of them when they are old. To give a sketch of that intuition, the following table shows the distribution of the responses to some questions asked about this expectation in the fourth round of the survey. The questions are worded in that way : To which extent do you expect the following kinds of help from [YL Child] when s/he is grown-up?

Table 1: *Percent of responses to some feeling questions (round 4)*

	Not at all	A little	Somewhat	Quite a lot	A lot	# obs
provides financial assistance to younger brothers/sisters	4.54	10.57	28.71	21.44	34.74	1,609
provides emotional support to you	1.56	4.09	12.76	19.92	61.66	1,857
helps you care for younger siblings	6.71	9.47	26.10	24.28	33.44	1,594
cares for you when you are old	2.90	5.84	13.57	21.29	56.41	1,865
provides financial assistance to you	4.17	9.42	25.79	22.42	38.20	1,869

More than half of caregivers expect child to care for them when they are old, and two third of caregivers expect emotional support from the child.

As a result, we consider as instruments[15] for parental investment the following variables considered as exogenous variations which affect cognitive and noncognitive skills only through their effect on parental investment :

- Shocks on household wealth, measured as the absolute variation of household wealth index between two periods.

- The child's birth order : an intuition is given by Behrman & Taubman (1986), who state : *If parents invest in children for insurance or for altruistic reasons that depend on the investment returns when the children become adults, it may be sensible to favor lower-order children since the financial or psychic returns are more likely to be available when the parents are still able to enjoy them*. In our data (round 4), we observe that 58% of caregivers expect the child to complete a graduate degree, and 18% of caregivers expect from child the completion of postgraduate degree. Their expectation could be an additional motivation to invest in the child.

[15]These instruments have also considered in Helmers & Patnam (2011).

Note that we do not distinguish the parental investment on cognitive skills and non-cognitive skills. We consider that they are the same, and this was also done in Cunha et al. (2010) and in Helmers & Patnam (2011) among others.

3 Data

3.1 Introduction

In order to answer our research question, we use data from the Ethiopia part of the *Young Lives* project, a long-term study of childhood poverty being carried out in 4 countries : India, Peru, Vietnam and of course Ethiopia. The broad objective of the *Young Lives* project is to improve understanding of the causes and consequences of childhood poverty and to examine how policies affect children's well-being. Extensive child, household and community level questionnaires are administered to capture information on various aspects of the child's life.

In Round 1, 1999 children aged around one (the "younger" cohort) and 1000 children aged around eight (the "older" cohort) were surveyed in 2002. Following up, Round 2 involved tracking the same children and surveying them in 2006 at age five and twelve respectively. Round 3 and 4 occurred in 2009 and 2014 respectively. Overall, 1866 children out of 1999 of the younger cohort have remained in the 4 waves of survey, which represents an attrition rate of 6.65% and 904 children out of 1000 of the older cohort remained in the four waves, which represents an attrition rate of 9.6%.

The sample of children is as follow : 5 regions (Addis Ababa, Amhara, Oromia, SNNP and Tigray) out of the 9 in Ethiopia were selected, accounting for 96 per cent of the national population. Then 3 to 5 districts were selected in each region with a balanced representation of food-deficient rural and urban districts. Then comes the choice of sentinel sites (20 overall selected); since districts were too large, in terms of both area and population, to be considered as sentinel sites, at least one peasant association per district was selected as a sentinel site, with the key criterion being the possibility of finding at least 100 households with a 1-year-old child and 50 households with an 8-year-old child. Then a village was randomly selected within each sentinel site. The questionnaires were then administered to around 100 one-year-old and 50 eight-year-old children in these villages. Data was collected through household questionnaires, child question-

naires and a community questionnaire. Our estimations incorporate this survey design, wherein we use the sentinel sites as our clustering variable.

We use data obtained from young cohort of children available in the currently four rounds of the Young Lives survey[16]. The two cohorts allow us to investigate two distinct periods of childhood. During the early childhood years, the transition between age one and five, a child still depends fully on her parents and family. The first few years of a child's life are decisive for the child's later physical and psychological well-being. The child learns during these years above all how to self-regulate, i.e., how to control his attention, emotions and behaviours. At the same time, the child acquires crucial cognitive skills, above all in terms of language acquisition. Therefore, the data on these early childhood years allow us to analyse factors influencing the foundations of skill formation, paying particular attention to a child's physical condition and his home environment.

3.2 Observed variables for latent factor estimation

The observed measurements used to estimate the latent factors are listed in the table 2 below.

The variables height-for-age Z-score, weight-for-age Z-score and BMI-for-age Z-score are continuous variables. They are used over the 4 rounds.

The observed variables used for latent cognitive factor are raw scores obtained in each of these tests. Indeed, the standardized Rasch score in the dataset contains so many missing values (until 75% of the inital sample would be missed if we considered Rasch scores); but as the raw score won't be used directly to estimate the technology of cognitive skills, this is not a drawback of our choice.

For cognitive skills at age 1, Cunha et al. (2010) used the child's weight at birth as a proxy for latent factor. They did not justify their choice and as a result we did not follow

[16] A fifth round is currently ongoing and data will be available by mid of 2018.

them and did not computer neither latent cognitive score, nor latent noncognitive score at age 1.

Table 2: *Observed variables used to construct the latent factors*

Round number	Health	Cognitive skills	Noncognitive skills	Parental investment
Round 1 (2002)	- Height-for-age Z-score - Weight-for-age Z-score			
Round 2 (2006)	- Height-for-age Z-score - Weight-for-age Z-score	- Peabody Picture Vocabulary Test (PPVT) - Cognitive Development Assessment test	- The child speaks and understands the commonly used language - The child travels to school with other children - The child does not feel in danger when travelling to school	- Share of parental expenditure on clothing for child - Share of parental expenditure on footwear for child
Round 3 (2009)	- Height-for-age Z-score - Weight-for-age Z-score	-Early Grade Reading Assessment test (EGRA) - Math test score - PPVT test score	- My friends look up to me as a leader - Do you find it hard to talk to other children? - If I try hard I can improve my situation in life - I think it is important to serve my community	- Share of parental expenditure on clothing for child - Share of parental expenditure on footwear for child - Share of parental expenditure on school fees for child - Share of parental expenditure on school books and stationery
Round 4 (2014)	- Height-for-age Z-score - BMI for age Z-score	- Language test score - Math test score - PPVT test score	- I make friends easily - If someone opposes me, I can find the means and ways to get ... - I'm as good as most other people - Overall, I have a lot to be proud of - I can always manage to solve difficult problems if I try hard	- Do you know YL Child's teacher? - A family member helps child with Homework

We will describe the variables used from round 2 to round 4 in the columns "Noncognitive skills" and "parental investment".

"The child speaks and understands the commonly used language" is based on a question on whether the child understand the commonly spoken language : three modalities are proposed : 0 for Not at all, 1 for Understands but does not speak and 2 for speaks and understands. Higher modalities imply a better fluency in the local language.

"The child travels to school with other children" is based on a question on how the child travel to school : 0 for alone, 1 for with parents or other adults and 2 for with other

children. Higher modalities express the degree of openness of the child to others.

"The child does not feel in danger when traveling to school" stands for a question on whether the child feels in danger when traveling to school, with modality 0 for Yes, child feels in danger and modality 1 for No, child does not feel in danger. A high modality indicates the courage of the child.

In round 3, the measurement of noncognitive skills are based on questions asking at which extent they agree with some statements. The answers are scaled from 1 (Strongly disagree) to 5 (Strongly agree), except for the statement "Do you find it hard to talk to other children?" where the answers are scaled from 1 (always) to 3 (Never). Overall, these questions capture how confident the child is.

In round 4, the measurement of noncognitive skills are based on questions asking at which extent they agree with some statements and the answers are coded from 1 (Strongly disagree) to 4 (Strongly agree). Higher modalities indicates a high self-esteem of the child.

For parental investment measurements, in rounds 2 and 3, the question asked for a given type of expenditure, what fraction has been devoted to the child selected for the survey. Note that several type of expenditures were included in the survey, from expenditures on school tuitions, books, transportation to school and health expenditures. But the huge amount of missing values leads us to selected only a few of them. This situation has an influence on the estimated latent factor of parental investment. In round four, two dichotomic variables have been used to describe how parents are concerned by the education of their children.

3.3 Missing data issues

The state space model used for the estimation of the latent factors does not allow for missing data. However, we did not use the methods advocated for handling missing data issues for two reasons. The first one is that I do not have much information on why

these data are missing. The second reason is that for some variables, we have up to 75% of missing values. The choices I did are the following

- For the child health indicators[17] at age 1, I imputed missing values by the mean of observed values. This is acceptable because only few of them were missing at age 1 and these variables are continuous;
- For all other indicators used as measurements for latent factors, I kept the ones with less missing values and I dropped missing values. I preserved the structure of the dataset in the sense that only children who have observed measures for proxies of latent factors have also values for the latent factors.

[17] Height for age and Weight for age z-scores

4 Results

We start by estimating a model of the state of the health of the children when they are 1 year old (precisely between 6 months and 17 months.

Table 3: *Estimation of child health model at age 1 (young cohort)*

	1-Health(SSM)	2-Health(PCA)	3-Serious illness
Level of antenatal neglect	-0.10+	-0.08*	-0.02
	[0.05]	[0.04]	[0.05]
No doctor present at birth	-0.36*	-0.28*	0.35*
	[0.14]	[0.10]	[0.16]
# months without breastfeeding	-0.07**	-0.06**	0.04**
	[0.01]	[0.01]	[0.01]
Unwanted pregnancy	-0.16	-0.14	0.08
	[0.12]	[0.09]	[0.08]
Caregiver experienced depression	-0.00	0.05	0.45**
	[0.10]	[0.08]	[0.08]
Highest grade caregiver completed	-0.03	-0.02	0.01
	[0.02]	[0.01]	[0.01]
wealth index	2.24**	1.52*	-1.35**
	[0.78]	[0.54]	[0.41]
Child sees daily biological dad	0.06	0.04	-0.11
	[0.11]	[0.08]	[0.09]
Child is male	-0.35**	-0.29**	0.14+
	[0.10]	[0.08]	[0.07]
Number of observations	1,469	1,469	1,468
Other controls added ?	Yes	Yes	Yes

+ $p < 0.1$; * $p < 0.05$; ** $p < 0.01$

Standard errors (in brackets) are clustered at sentinel level. Column 3 is a probit estimation

We built a latent health indicator by using two variables : the weight for age and the height for age. These indicators are proposed by the World Health Organization as child growth standards. Higher values are an expression of a better health. These indicators are embodied in a state space model with one state equation and two measurement equations. The latent health indicator is the smooth estimate of the state variable of the model.

We relate the health indicator of a child at age 1 to his antenatal conditions, early childhood conditions and household characteristics. The results above show that the

household wealth and lack of breastfeeding[18] are the most important determinants of the health of the child. This finding is robust to the use of different methods to estimate the latent health variable indicator; indeed, the latent factor used in column 1 is obtained by a state space representation of z-scores, and column 2 is obtained by Principal component analysis of the two z-scores.

Column 3 is a probit estimation based on a binary variable on whether or not the child has suffered of a serious illness so that the mother thought that he would die. A higher wealth index decrease the probability that the child suffers of such a serious illness, while the absence of a doctor at birth, the number of months without breastfeeding and the fact that the caregiver experienced a depression increase significantly the probability for the child to suffer of a serious illness.

The results show also that male children have a poor health compared to female children. We test on each z-score the equality of means and obtained a statistically significant difference of height for age z-score and weight for age z-score between male and female, female having a higher z-score. See for illustration the graph n°1 in the appendix.

The next table (n°4) presents the results of the relation between Cognitive skills and child health/parental investment. Child health and household wealth have a positive and statistically significant effect on cognitive skills. Because we poorly measured parental investment due to huge missing data (we only use expenditures on footwear and clothing which are not directly linked to investment on cognitive ability), we have a negative and significant effect of parental investment.

Now we estimate our first transition of skill building, from age 5 to age 7 (see table n°5.

[18]We computed this variable as the number of months the child is left without breastfeeding compared to a standard of 16 months, as done in Helmers & Patnam (2011).

Table 4: *Estimation of cognitive skills model at age* 5 *(young cohort)*

	1-CS	2-CS	3-CS	4-Inv	5-CS(2SLS)
Child's parental investment (Age 5)	0.14	0.15	0.14		-3.66*
	[0.10]	[0.10]	[0.10]		[1.68]
Child's health (age 1)	0.04	0.11*			
	[0.05]	[0.04]			
Child's health (age 5)	0.14*		0.16**	0.01	0.22**
	[0.05]		[0.04]	[0.01]	[0.06]
Caregiver's education level	0.19**	0.19**	0.19**	0.01	0.24**
	[0.04]	[0.04]	[0.04]	[0.01]	[0.06]
Household size	0.06+	0.06+	0.06+	-0.16**	-0.54*
	[0.03]	[0.03]	[0.03]	[0.02]	[0.25]
wealth index	1.96**	2.09**	2.01**	0.15	2.46**
	[0.66]	[0.70]	[0.66]	[0.32]	[0.88]
Gender	0.20	0.22+	0.18	0.04	0.31
	[0.12]	[0.12]	[0.12]	[0.04]	[0.19]
Age of child in months	0.15**	0.15**	0.15**	0.00	0.16**
	[0.02]	[0.02]	[0.02]	[0.01]	[0.03]
Urban area	0.87*	0.85*	0.87*	0.16	1.37+
	[0.31]	[0.33]	[0.31]	[0.16]	[0.80]
Number of siblings	-0.02	-0.02	-0.03	0.03	-0.32+
	[0.04]	[0.05]	[0.04]	[0.04]	[0.18]
CH birth order				-0.11**	
				[0.04]	
Shock on household wealth				0.07	
				[0.38]	
Number of observations	1,364	1,364	1,364	1,392	1,353
P-value for endogeneity test					.004

+ $p < 0.1$; * $p < 0.05$; ** $p < 0.01$

Parental investment is instrumented by child's birth order and shocks on HH wealth. Standard errors (in brackets) are clustered at sentinel level

The results show self-productivity effect of cognitive skills and noncognitive skills, but no cross-productivity effect. In addition, parental investment at age 5 does not increase neither cognitive skills, nor noncognitive skills at age 7. The explanation is the weak measure of parental investment used at age 5, because of missing data issues.

Table 5: *Estimation of cognitive and noncognitive skills models at age 7 (young cohort)*

	1-CS	2-NCS	3-Inv	4-CS(2SLS)	5-NCS(2SLS)
Child's parental investment (Age 5)	-0.04	0.00		0.10	0.00
	[0.06]	[0.00]		[0.09]	[0.01]
Cognitive skills (age 5)	0.07**	0.00	0.01	0.07**	0.00
	[0.01]	[0.00]	[0.01]	[0.01]	[0.00]
Noncognitive skills (age 5)	-0.33	0.02+	-0.21**	-0.29	0.02+
	[0.20]	[0.01]	[0.07]	[0.19]	[0.01]
Child's health (age 7)	0.29**	0.00	0.03	0.28**	0.00
	[0.09]	[0.00]	[0.04]	[0.09]	[0.00]
Caregiver's education level	0.03*	-0.00+	0.01	0.02*	-0.00+
	[0.01]	[0.00]	[0.01]	[0.01]	[0.00]
Household size	-0.03	0.00	-0.09**	-0.00	0.00
	[0.03]	[0.00]	[0.01]	[0.03]	[0.00]
wealth index	1.48**	0.05*	-0.19	1.51**	0.05*
	[0.46]	[0.02]	[0.24]	[0.46]	[0.02]
Gender	0.04	0.00	0.05	0.04	0.00
	[0.06]	[0.00]	[0.05]	[0.06]	[0.00]
Age in months	0.05**	0.00	-0.00	0.05**	0.00
	[0.01]	[0.00]	[0.01]	[0.01]	[0.00]
Urban or Rural	-0.73+	-0.03	0.00	-0.75*	-0.03
	[0.40]	[0.02]	[0.11]	[0.38]	[0.02]
CH birth order			-0.13**		
			[0.01]		
Shock on household wealth			-0.15		
			[0.31]		
Number of observations	1,282	1,286	1,344	1,275	1,279
R^2	0.51	0.05	0.28	0.50	0.05
P-value for endogeneity test				.123	.983

+ $p < 0.1$; * $p < 0.05$; ** $p < 0.01$

Parental investment is instrumented by child's birth order and shocks on HH wealth. Standard errors (in brackets) are clustered at sentinel level

Table 6: *Estimation of cognitive and noncognitive skills models at age* 12 *(young cohort)*

	1-CS	2-NCS	3-Inv	4-CS(2SLS)	5-NCS(2SLS)
Parental Investment (age 7)	0.29+	0.00		0.66	-0.00
	[0.16]	[0.00]		[0.46]	[0.02]
Cognitive skills (age 7)	3.42**	0.03*	0.07	3.42**	0.03**
	[0.74]	[0.01]	[0.07]	[0.73]	[0.01]
Noncognitive skills (age 7)	10.14	-0.32	-0.50	10.57	-0.31
	[9.95]	[0.23]	[0.77]	[9.78]	[0.22]
CH health (age 12)	10.72**	-0.05**	0.33**	10.56**	-0.05**
	[1.27]	[0.02]	[0.11]	[1.18]	[0.02]
Caregiver's education level	-0.02	0.01	0.01	-0.03	0.01
	[0.13]	[0.00]	[0.02]	[0.13]	[0.00]
Household size	-1.10**	0.01	-0.28**	-0.96**	0.01
	[0.32]	[0.01]	[0.03]	[0.37]	[0.01]
Wealth index	19.77**	0.00	-0.40	19.86**	-0.00
	[5.05]	[0.10]	[0.59]	[4.95]	[0.10]
Sex of YL Child	-0.93	0.03	0.07	-0.95	0.04
	[0.77]	[0.03]	[0.11]	[0.74]	[0.03]
Child age	-0.06	-0.00	0.00	-0.06	-0.00
	[0.10]	[0.00]	[0.02]	[0.10]	[0.00]
CH birth order			-0.30**		
			[0.03]		
Shock on Household wealth			0.17		
			[0.53]		
Number of observations	1,000	1,302	1,306	996	1,298
P-value for endogeneity test				.428	.767

$+\ p < 0.1$; * $p < 0.05$; ** $p < 0.01$
Parental investment is instrumented by child's birth order and shocks on HH wealth. Standard errors (in brackets) are clustered at sentinel level

The second transition of skills building (table n°6) show evidence of self-productivity for cognitive skills and cross-productivity from cognitive skills to noncognitive skills. This finding has also been obtained by Helmers & Patnam (2011) using the first and second rounds for the older cohort of the Young Lives survey in India. Still, we don't see any effect of parental investment, although an effort has been made to include expenditures related to books and stationery. We should also note a negative effect of child health on noncognitive skills at age 12.

5 Conclusion

There is an agreement among researchers that ability explains a substantial part of the differences across people of success in socioeconomic life, and that ability gaps across people emerge at childhood before they start school. Building on recent advances in the child development literature in Economics pioneered by James J. Heckman, we estimate two transitions technology of skill formation using four waves of survey data in Ethiopia which are part of the longitudinal project "Young lives" funded by the UK aid department.

We found evidence that early life conditions, including antenatal care have significant effects on child' health, and that child health is positively related to a higher level of ability in the four round, except at age 12 where we observed a negative effect of child health on noncognitive skills. We also found evidence of self-productivity for cognitive skills and noncognitive skills and cross productivity from cognitive to noncognitive skills.

The lack of effect of parental investment on either cognitive skills, or noncognitive skills is due to missing data issues which made us unable to use some important measures of parental investment. We cannot thus conclude about the trade-off between early and later investment on disadvantaged children.

References

Behrman, J. R. & Taubman, P. (1986), 'Birth Order, Schooling, and Earnings', *Journal of Labor Economics* 4(3), $S121$–$S145$.
URL: *http://www.jstor.org/stable/2534958*

Bernard, T. & Taffesse, A. S. (2014), 'Aspirations: an approach to measurement with validation using ethiopian data', *Journal of African Economies* 23(2), 189–224.

Blume, L. E., Brock, W. A., Durlauf, S. N. & Ioannides, Y. M. (2010), Identification of social interactions, *in* J. Benhabib, A. Bisin & M. O. Jackson, eds, 'Handbook of Social Economics, Volume 1B', North Holland, pp. 853–964.

Bono, E. D., Francesconi, M., Kelly, Y. & Sacker, A. (2016), 'Early Maternal Time Investment and Early Child Outcomes', *The Economic Journal* 126(596), F96–F135.
URL: *http://onlinelibrary.wiley.com/doi/10.1111/ecoj.12342/abstract*

Card, D. (1999), Chapter 30 - The Causal Effect of Education on Earnings, *in* O. C. A. a. D. Card, ed., 'Handbook of Labor Economics', Vol. 3, Part A, Elsevier, pp. 1801–1863.
URL: *http://www.sciencedirect.com/science/article/pii/S1573446399030114*

Coneus, K., Laucht, M. & Reuß, K. (2012), 'The role of parental investments for cognitive and noncognitive skill formation—Evidence for the first 11 years of life', *Economics & Human Biology* 10(2), 189–209.
URL: *http://www.sciencedirect.com/science/article/pii/S1570677X11000062*

Cunha, F. & Heckman, J. J. (2007), 'The technology of skill formation', *American Economic Review* 92(2), 31–47.

Cunha, F. & Heckman, J. J. (2008), 'Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation', *The Journal of Human Resources* 43(4), 738–782.

Cunha, F., Heckman, J. J. & Schennach, S. M. (2010), 'Estimating The Technology of Cognitive and Noncognitive Skill Formation', *Econometrica* 78(3), 883 − −931.
URL: *http://www.jstor.org/stable/40664515*

de Haan, M. (2011), 'The effects of parents' schooling on child's schooling : A nonparametric bound analysis', *Journal of Labor Economics* 29(4), 859–892.

De Jong, P. (1991), 'The diffuse kalman filter', *Annals of Statistics* 19, 1073–1083.

Durbin, J. & Koopman, S. J. (2012), *Time series analysis by state space methods*, Oxford University Press. Second edition.

Figlio, D., Guryan, J., Karbownik, K. & Roth, J. (2014), 'The effects of poor neonatal health on children's cognitive development', *The American Economic Review* 104(12), 3921–3955.
URL: *http://www.jstor.org/stable/43495361*

Gourieroux, C. & Monfort, A. (1995), *Statistics and econometric models*, Cambridge University Press. volume 1 : general concepts, estimation, prediction and algorithms.

Heckman, J. J. (2008), 'Role of income and family influence on child outcomes', *Annals of the New York Academy of Sciences* 1136(1), 307–323.
URL: *http://dx.doi.org/10.1196/annals.1425.031*

Helmers, C. & Patnam, M. (2011), 'The formation and evolution of childhood skill acquisition: Evidence from india', *Journal of Development Economics* 95(2), 252 – 266.
URL: *http://www.sciencedirect.com/science/article/pii/S0304387810000295*

Holmlund, H., Lindahl, M. & Plug, E. (2011), 'The causal effect of parents' schooling on childrens's schooling : A comparison of estimation methods', *Journal of Economic Literature* 49(3), 615–651.

Hunter, P. (2008), 'What genes remember', *Prospect* (146). Last accessed on May 19th 2017.
URL: *https://www.prospectmagazine.co.uk/magazine/whatgenesremember*

Jacob, B. & Rothstein, J. (2016), 'The measurement of student ability in modern assessment systems', *The Journal of Economic Perspectives* 30(3), 85–107.
URL: *http://www.jstor.org/stable/43855702*

Koch, A., Nafziger, J. & Nielsen, H. S. (2015), 'Behavioral economics of education', *Journal of Economic Behavior & Organization* 115, 3–17.
URL: *https://www.sciencedirect.com/science/article/pii/S0167268114002431*

Manski, C. F. & Pepper, J. V. (2000), 'Monotone instrumental variables : with an application to the returns to schooling', *Econometrica* 68(4), 997–1010.

Manski, C. F. & Pepper, J. V. (2009), 'More on monotone instrumental variables', *Econometrics Journal* 12, 200–216.

Thuilliez, J., Sissoko, M. S., Toure, O. B., Kamate, P., Berthelemy, J. C. & Doumbo, O. K. (2010), 'Malaria and primary education in mali: a longitudinal study in the village of donéguébougou', *Soc Sci Med.* 71(2), 324–334.

Todd, P. E. & Wolpin, K. I. (2007), 'The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps', *Journal of Human Capital* 1(1), 91–136.

6 Appendix

Figure 1: *Comparison of z-scores between female and male children at age* 1

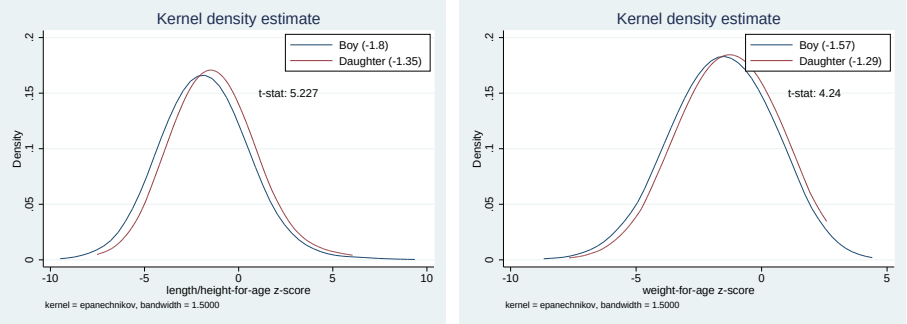

Table 7: *Descriptive statistics : round* 1 *to* 4

	Round 1	Round 2	Round 3	Round 4
Child is male	.53	.53	.53	.53
	(1999)	(1912)	(1884)	(1872)
Child age	12	62	97	145
	(1999)	(1912)	(1884)	(1871)
CH birth order		3.4		
	(0)	(1907)	(0)	(0)
Able to speak & understand lang		1.9	1	
	(0)	(1912)	(1886)	(0)
Raw score in Math test			6.6	11
	(0)	(0)	(1808)	(1623)
Raw score in PPVT test		21	79	38
	(0)	(1861)	(1857)	(1640)
Raw score in CDA test		8.2		
	(0)	(1888)	(0)	(0)
Raw score in EGRA test			5.1	
	(0)	(0)	(1879)	(0)
Height for age z-score	-1.6	-1.5	-1.2	-1.5
	(1999)	(1909)	(1882)	(1871)
Weight for age z-score	-1.4	-1.4	-1.6	
	(1999)	(1909)	(1882)	(0)
BMI for age z-score				-1.8
	(0)	(0)	(0)	(1870)
Household size	5.7	6	6.2	5.9
	(1999)	(1912)	(1886)	(1874)
Wealth index	.21	.28	.33	.37
	(1977)	(1902)	(1885)	(1871)
Urban area	.35	.4	.4	
	(1999)	(1912)	(1886)	(0)
CH born at home	.82			
	(1995)	(0)	(0)	(0)
No doctor present at birth	.89			
	(1774)	(0)	(0)	(0)
Unwanted pregnancy	.38			
	(1894)	(0)	(0)	(0)
Breastfeeding privation	-4.7			
	(1895)	(0)	(0)	(0)
Antenatal neglect	2.1			
	(1847)	(0)	(0)	(0)
Health	-1.9	-3.2	-.00008	-.1
	(1999)	(1909)	(1882)	(1870)
Cognitive skills		6.1	3.3	140
	(0)	(1860)	(1787)	(1423)
Noncognitive skills		3	1.6	4.9
	(0)	(1895)	(1801)	(1856)
Parental investment		2	4.9	3.3
	(0)	(1564)	(1621)	(1811)

Figures in () are the number of observations.