FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ECONOMIA DE SÃO PAULO

LUCIANA NEVES PIRES

THE IMPACTS OF EARLY CHILDHOOD INVESTMENT: AN APPROACH THROUGH LATENT COGNITIVE SKILLS

SÃO PAULO
2018

LUCIANA NEVES PIRES

THE IMPACTS OF EARLY CHILDHOOD INVESTMENT: AN APPROACH THROUGH LATENT COGNITIVE SKILLS

Dissertação apresentada à Escola de Economia de São Paulo da Fundação Getulio Vargas como requisito para obtenção do título de Mestre em Economia de Empresas

Campo de Conhecimento:
Microeconomia Aplicada

Orientador: Cristine Campos de Xavier Pinto

São Paulo

2018

Pires, Luciana Neves.
The impacts of early childhood investment: an approach through latent cognitive skills / Luciana Neves Pires. - 2018.
99 f.

Orientador(a): Cristine Campos de Xavier Pinto.
Dissertação (CMEE) - Escola de Administração de Empresas de São Paulo.

1. Educação pré-escolar. 2. Educação de crianças. 3. Crianças - Desenvolvimento. 4. Investimentos na educação. I. Pinto, Cristine Campos de Xavier. II. Dissertação (CMEE) - Escola de Administração de Empresas de São Paulo. III. Título.

CDU 37.046.12

Ficha catalográfica elaborada por: Raphael Figueiredo Xavier CRB SP-009987/O
Biblioteca Karl A. Boedecker da Fundação Getulio Vargas - SP

LUCIANA NEVES PIRES

THE IMPACTS OF EARLY CHILDHOOD INVESTMENT: AN APPROACH THROUGH LATENT COGNITIVE SKILLS

Dissertação apresentada à Escola de Economia de São Paulo da Fundação Getulio Vargas como requisito para obtenção do título de Mestre em Economia de Empresas.

Campo de Conhecimento:
Microeconomia Aplicada

Data de Aprovação:
___/___/_____

Banca examinadora:

_____
Prof. Dr. André Portela Souza
FGV-EESP

_____
Profa. Dra. Cristine Campos de Xavier Pinto
FGV-EESP

_____
Prof. Dr. Daniel Domingues dos Santos
FEA-RP/USP

_____
Prof. Dr. Vladimir Pinheiro Ponczek
FGV-EESP

AGRADECIMENTOS

Aos meus pais, Geraldo e Maria Aparecida, por sempre investirem e incentivarem os estudos, e ao Marcos, pelo especial apoio nessa etapa.

Ao meu namorado José Carlos, que me apoiou tão de perto em cada passo durante esse período.

À professora Cristine, pela orientação, disponibilidade e suporte na elaboração desse trabalho.

Ao professor Daniel Santos e à equipe do LEPES, pela atenção e tempo dispensados, e pelo fornecimento de dados cruciais para a elaboração desse trabalho.

Aos professores Vladimir Ponczek e André Portela, pela disponibilidade em participar da banca e pelos valiosos comentários sobre o trabalho.

Ao Luis, pela paciência e pela inestimável ajuda com os códigos.

À Gabriela e à Natália, sempre solícitas para tirar dúvidas sobre a base de dados.

Aos colegas de turma, que tornaram essa jornada possível.

ABSTRACT

This work contributes to the literature of human capital formation by estimating the impact of early childhood investment (by means of preschool attendance) over cognitive skill development. Using a longitudinal panel dataset for a single municipality, we draw the distribution of latent cognitive and non-cognitive factors and consider a dynamic model of skill formation. We find our Constant Elasticity Substitution (CES) production function to be a Cobb-Douglas. The complementarity and share parameters of the CES are stable in diverse specifications tested. We find that early investment matters for cognitive skill accumulation during childhood. Preliminar estimation for long-term cognitive production presents evidence of self-productivity (skill begets skill). Since cognitive skill is persistent overtime, early childhood investment has a positive cumulative impact in the long-term by boosting cognitive skills in earlier stages.

Key-words: Preschool, Early Childhood Investment, Cognitive Skills, Socioemotional Skills, Skill Production Function.

RESUMO

Este trabalho contribui para a literatura de formação de capital humano estimando o impacto do investimento na primeira infância (por meio da frequência pré-escolar) sobre o desenvolvimento de habilidades cognitivas. Usando um painel longitudinal com dados para um único município, extraímos a distribuição dos fatores cognitivo e não-cognitivo e consideramos um modelo dinâmico de formação de habilidades. A função de produção Elasticidade Substituição Constante (CES) que estimamos é uma Cobb-Douglas. Os parâmetros de complementaridade e participação na CES são estáveis para diversas especificações testadas. Encontramos que o investimento na primeira infância é importante para o acúmulo de habilidades cognitivas durante a infância. Estimativas preliminares da função de produção usando medidas de longo prazo para o fator cognitivo apresentam evidências de autoprodutividade (habilidade gera habilidade). Como a habilidade cognitiva é persistente ao longo do tempo, o investimento na primeira infância tem um efeito cumulativo positivo no longo prazo, impulsionando as habilidades cognitivas nos estágios iniciais.

Palavras-chaves: Ensino Infantil, Investimento na Primeira Infância, Habilidades Cognitivas, Habilidades Socioemocionais, Função de Produção de Habilidade.

List of Figures

List of Tables

Table of Contents

1 Introduction

Cognitive and socioemotional skills are shown to be highly predictive about the degree of socioeconomic success achieved by individuals. There is evidence that acquired skills influence schooling level, academic performance, behavior and decision making, labor market outcomes, in sum, influencing one's ability to fully develop their potential and interact in society (HECKMAN; STIXRUD; URZUA, 2006).

The idea that skills matter, combined with the fact that they are produced and somewhat malleable, motivated a number of works dedicated to investigate the dynamics of skill production throughout the lifespan (CUNHA; HECKMAN, 2007; HECKMAN, 2006). There is substantial evidence that skills are affected by environment stimuli, parental investments and the previous level of skills, in a complex dynamic interplay. Evidence that skills are produced in a cumulative fashion, with the existence of sensible and crucial periods for development and different productivity rates at different stages, raises a discussion on the optimal investment pattern throughout childhood.

In particular, the first years of life are a crucial period for brain development (PHILLIPS; SHONKOFF et al., 2000). From the perspective of efficiency, the early childhood would be key because investments made in this period would have higher returns (in comparison to late investments) due to the cumulative effect over the following periods (CUNHA; HECKMAN, 2007; HECKMAN, 2006). From the perspective of equality, early interventions could contribute to reduce the skill gap among children from vulnerable and favorable backgrounds.

Small scale programs in developed countries, such as Perry Preschool and Carolina Abecederian, are estimated to have long lasting effects over cognitive abilities and labour market outcomes (HECKMAN et al., 2010; CAMPBELL et al., 2002), while the effects of large scale programs are estimated to fade out as long as time passes.

There is a growing number of studies seeking to evaluate the effects of early childhood investment in cognitive skill for disadvantaged children in the context of developing countries, such as Attanasio et al. (2015) and Attanasio, Meghir and Nix (2015). For the case of Brazil, previous works found that preschool attendance is associated to better academic performance (as measured by standardized test scores), higher chances of completing mandatory education and higher wages. One of the greatest challenges of this literature is to properly compare preschool attendees and non-attendees, since parental decision of when to enroll a child in school is not random.

This work aims to contribute to the literature of human capital formation by

estimating the impact of early childhood investment (by means of preschool attendance) over cognitive skill development, using a dynamic model as of Cunha, Heckman and Schennach (2010). We draw cognitive and non-cognitive latent factors using longitudinal panel data for children in Sertãozinho/SP. We provide an instrument to address the endogeneity of parental decision on early childhood investment, and use a control function approach to include it as an input for cognitive skill production. We find that preschool attendance does matter for cognition. In line with the current literature, we find that our Constant Elasticity Substitution (CES) production function is a Cobb-Douglas. We also present preliminary results for long-term indirect effects of preschool attendance over cognitive skills.

This work is divided as follows. Chapter 2 contains a brief review of previous works, and how does the present work is placed in the existing literature. In chapter 3 we present the theoretical model for cognitive skill formation with endogenous parental investment decision. Our database is presented in chapter 4. In chapter 5 we expose the model adapted to our data and the estimation strategy used. Finally, results are reported and discussed in chapter 6.

2 Literature Review

Latent skills are vastly documented to explain labor market outcomes. Not only cognitive skills, but also non-cognitive or socioemotional skills [1] greatly matter to explain schooling achievement, income and other outcomes of interest for society (HECKMAN; STIXRUD; URZUA, 2006; ACOSTA; MULLER; SARZOSA, 2015).

Cunha and Heckman (2007) develop a model of skill formation throughout childhood. A flexible theoretical framework embodies some of the findings in the empirical works, such as the existing evidence of self-productivity (skill begets skill) and of cross productivity (a skill contributing to the accumulation of another skill). Using a dynamic production function, they discuss what could be the optimal investment path throughout childhood. The early childhood is likely to be a crucial period since early investments would have more time to be harvested, and would have a cumulative impact over the whole development path.

Cunha, Heckman and Schennach (2010) generalize previous findings in the literature by proposing a dynamic model of skill formation. The linearity assumption in the skill production function [2] is relaxed by allowing the skill production function to be a CES. Inputs are allowed to freely correlate and interact to produce outputs. The technology of skill formation is identified nonparametrically building on Carneiro, Hansen and Heckman (2003), and estimated using nonlinear factor models.

A substantial part of the evidence in favor of early childhood programs comes from high quality and intensive programs in the developed world, such as Perry Preschool and Carolina Abecederian Programs (HECKMAN et al., 2010; CAMPBELL et al., 2002). Apart from those high quality programs, the estimated impacts of preschool attendance seems to be over cognitive rather than non-cognitive outcomes (VERAMENDI; URZÚA, 2011).

There is a growing number of works investigating whether daycare and preschool enrollment impacts are positive, in particular in the context of developing countries. The main challenge of this literature is to address the endogeneity of parental investment decision and the sources of unobserved heterogeneity, in order to obtain a causal effect of early childhood investment over child's skills and outcomes.

Veramendi and Urzúa (2011) investigate the effects of daycare attendance using data for Chile. The endogeneity of daycare is addressed with instruments, and unobserved heterogeneity is treated using the approach of latent abilities. A structural model is

[1] Both terms are used interchangeably in this work.
[2] As of assumed in Cunha and Heckman (2008)

estimated using Maximum Likelihood estimation. First the distribution of mother and child latent abilities is drawn from a dedicated linear measurement system, such that each group of measures is only affected by the corresponding latent factor. Then the models for the daycare enrollment decision and child's outcomes are run using the estimates from the first stage. Separate models are run for children that attended daycare and the ones that did not. Evidence is found that children from high ability mothers are more likely to have attended daycare, and that child care center enrollment would boost cognitive development, even after controlling for selection.

Attanasio, Meghir and Nix (2015) investigate the importance of investments in skill formation by estimating production functions for two dimensions of human capital, cognition and health, in the context of India. Following Cunha, Heckman and Schennach (2010), endogenous investments throughout childhood are allowed to respond to unobserved shocks affecting human capital formation. Endogeneity is addressed using variation on the prices of investment goods as instruments,in a sort of a first stage, and the residual of the investment equation is included as a control in the production functions for cognition and health. Investment is found to affect the production of both latent factors during childhood, with a higher marginal product for higher levels of both cognition and health.

Building on the same theoretical and empirical framework of Attanasio, Meghir and Nix (2015), Attanasio et al. (2015) and Attanasio et al. (2017) findings are consistent with the preceding literature of skill production. Cognition and health are found to be persistent overtime in a complex interplay, being health determinant for cognition development and parental investment important for both. Moreover, Attanasio et al. (2017) call attention to the importance of allowing for flexibility in the functional form of production function, since the nested CES performed better results in several cases, using data for children aged 1-15 years old in Peru and Ethiopia. Attanasio et al. (2015) evaluate how a randomized intervention in Colombia could have changed parental time and material investment in children aged 18-36 months. They found evidence of strong complementarity among parental investment, parents' skills and child's previous skills.

The current literature on the effect of preschool attendance using Brazilian data in general considers as outcome the academic performance as measured by large scale standardized test scores, as the Basic Education National Evaluation System (SAEB and Prova Brasil scores) [3]. Posing a linear measurement system to draw latent abilities, as of Cunha, Heckman and Schennach (2010), and investigating the effect of early investments on skills - rather than on proficiency scores - requires longitudinal panel data, which is not

[3] Both exams are complementary in the Basic Education National Evaluation System, and are usually mentioned followed by the corresponding year. The SAEB evaluates samples of students from public and private schools. When the exam application is censitary, for Portuguese and Math Exams for students in the 5^{th} and 9^{th} years of elementary school, the exam is called Prova Brasil. Further information in http://portal.inep.gov.br/educacao-basica/saeb and http://portal.mec.gov.br/prova-brasil

usually available for such large samples.

Using standard econometric methods, preschool attendance is estimated to have a positive effect over proficiency scores, wages and the probability of being employed, even after controlling for race, gender, parents' schooling, household income, among other controls (CURI; MENEZES-FILHO et al., 2006; FELÍCIO; MENEZES; ZOGHBI, 2010; FELÍCIO; FERNANDES, 2005; FELÍCIO; VASCONCELLOS et al., 2007).

Curi, Menezes-Filho et al. (2006) obtained a positive impact of attending preschool on the probability of concluding school cycles and on wages, using logit and Ordinary Least Squares (OLS) regressions over cohorts in a survey from IBGE [4]. Preschool attendance was found to be associate to higher wages regardless of the schooling level achieved, in addition to a positive impact over proficiency scores SAEB 2003. Indeed, Felício and Fernandes (2005) estimated that preschool attendance accounted for 28% of the standard deviation of Math scores (an average effect calculated for schools that participated of SAEB 2003) and of 39% of the standard deviation national average SAEB and Prova Brasil 2005 scores.

Pinto, Santos and Guimarães (2017) found daycare and kindergarten attendance to impact Math scores in the 4^{th} year of elementary education by 0.22 standard deviations, as measured by SAEB 2005. They make an attempt to disentangle parental decision on daycare and kindergarten enrollment using a control function approach. The endogeneity of parents' choice is instrumented with the supply of daycare and kindergarten in the municipality level, and a control function is used to correct for selection effects. Daycare and kindergarten are found to have positive and heterogeneous effects on child proficiency, being the impact on scores increasing in mother's education.

The present work also uses a control function approach and a measure of supply of daycare/kindergarten as instrument to endogenous parental decision on preschool attendance, as of Pinto, Santos and Guimarães (2017). However, our empirical strategy relies on Cunha, Heckman and Schennach (2010) and Attanasio, Meghir and Nix (2015). In addition, our database consist of a single municipality in the countryside of São Paulo state, Sertãozinho.

Using the first round of Sertãozinho field survey, with data collected for the year of 2008, Felício, Terra and Zoghbi (2012) found that children that attended preschool had a better performance in Provinha Brasil exam, relative to the appropriate controls (assigned with Propensity Score Matching) that started school in mandatory education only. Working in the same data, Fonseca (2015) used the balanced longitudinal panel for Sertãozinho that had information collected in 2008 and 2012 years. The purpose was to investigate whether the perceived benefits of attending daycare and kindergarten would decrease overtime (the fading out effect). Using selection on observables and a couple

[4] Standard of living survey - Pesquisa de Padrão de Vida (PPV) - was conducted by the Brazilian Institute of Geography and Statistics (IBGE) in the years of 1996-1997.

more exercises using OLS specifications, Fonseca (2015) makes an effort to compare how preschool attendees and non attendees had performed in proficiency scores in 2012 relative to their performance in 2008. Although the hypothesis of fading out could not be fully rejected, the effect of preschool attendance on test scores was estimated to be positive and persistent.

Our results are consistent with previous work in the same database and with preceding findings in this literature. We do find a positive impact of preschool enrollment over cognitive skill formation, measured by Provinha and proficiency scores, even after controlling for observables that could be affecting child development, as Felício, Terra and Zoghbi (2012). We estimate early childhood investment to be complementary to socioemotional skills for cognitive skill production, and cognitive skills to be persistent from childhood to teenage years. Moreover, if it is the case that the positive impact of preschool fades out, we argue that there is a long lasting effect due to its impact on early cognition.

Procópio et al. (2012) considered a dynamic production function for cognitive skills, considering the proficiency as ability itself, for a cohort of students accompanied from 2005 to 2008 in GERES longitudinal data. The impact of early childhood inputs is measured using a dynamic skill production function, adapted from Todd and Wolpin (2003), where each period's proficiency level in Portuguese and Mathematics is a linear function of previous periods' proficiency interacted with previous inputs. Using OLS and accounting for nonobserved heterogeneity, it was found that that inputs along childhood have a cumulative effect over school outcomes, as measured by standardized test scores. Results also point that preschool investment could contribute to reduce inequality in math proficiency.

We do not impose the human capital accumulation to be linear (as of Todd and Wolpin (2003) and do not rely on proficiency scores as directly measuring skills, as the majority of previous works that evaluate preschool attendance effects in Brazil. Our work lays on the same theoretical ground as Ulyssea (2017), using a dynamic model for skill formation, building on Cunha, Heckman and Schennach (2010), and following the empirical strategy of Attanasio, Meghir and Nix (2015). Using a cohort of Brazilian children, Ulyssea (2017) find that parental investment plays an important role in cognitive and socioemotional skills formation from gestation to age 18. Child previous skill level, parental health and abilities were found to be complementary in the human capital formation process. A structural model is also used to evaluate how a national Chilean program impacted children's ability and health. As of Attanasio et al. (2015), it was possible to disentangle effects in the magnitude of the latent variables and on the parameters in the production functions. The program was effective in changing parental investments and would have contributed to reduce inequality by affecting higher disadvantadged children.

3 Model

The purpose of this work is to investigate the impact of early childhood investments (by means of daycare and kindergarten enrollment) on child's skill formation. We consider that the child i's latent skill has two dimensions, $\theta_i = (\theta_i^C, \theta_i^N)$, where θ_i^C represents cognitive skills and θ_i^N non-cognitive skills. The model for cognitive skill accumulation builds on Cunha, Heckman and Schennach (2010) and Attanasio, Meghir and Nix (2015).

Each child i is endowed with $\theta_{i,0}$ at birth, a latent factor that reflects intrinsic ability, pre-birth environment, *in-utero* experiences. We consider $t \in \{1, \ldots, T\}$ periods in childhood, such that period $t = 1$ represents early childhood, when the child is aged 0-6 years old.

In each period, latent skill vector is given by $\theta_{i,t} = (\theta_{i,t}^C, \theta_{i,t}^N)$. Parents do not directly observe $\theta_{i,t}$, but the child human capital $h_{i,t}$ in each period, being

$$h_{i,0} \equiv \theta_{i,0} \tag{3.1}$$

$$h_{i,t} = g(\theta_{i,t}^C, \theta_{i,t}^N) \tag{3.2}$$

We assume the function g is continuous, twice differentiable, increasing in both arguments and bounded.

3.1 Cognitive skill production function

We provide a model for the dynamics of cognitive skill accumulation. Non-cognitive skills are assumed to be stable during childhood, due to the absence of available data. Even though this is an assumption we are not able to relax, it might not be such a strong one.

There is a discussion in the neuroscience and psychology literature of whether psychometric tests - as of the well-know Big Five Inventory (BFI) Questionnaire - would reflect personality traits, stable in the long-term, or abilities, malleable and subject to change overtime. There is a branch in the literature that supports the first hypothesis, finding high persistence of personality traits during adulthood (COBB-CLARK; SCHURER, 2012; RANTANEN et al., 2007; SOLDZ; VAILLANT, 1999), or even during the life cycle (BORGHANS et al., 2008). The substitution of inputs in the production of non-cognitive skills is found to be nearly constant during different stages in the life cycle (CUNHA; HECKMAN; SCHENNACH, 2010).

Therefore, we argue it is acceptable to consider that, despite socioemotional skills can change during a lifetime, due to intrinsic maturation and/or interacting with other abilities and environment factors, they can be considered to remain stable during childhood. For simplicity of notation, we refer to non-cognitive skill as θ_i^N, without the time subscript, and represent the latent stock of ability in every period $t \geq 1$ as $\theta_{i,t} = (\theta_{i,t}^C, \theta_i^N)$.

The dynamics of cognitive skill production is given as follows,

$$\theta_{i,t+1}^C = f(\theta_{i,t}^C, \theta_i^N, I_{i,t}, \eta_{i,t}) \tag{3.3}$$

where $I_{i,t}$ accounts for parental investment in period t and $\eta_{i,t}$ are shocks affecting cognitive skill's production. In particular, we let f be a CES function [1], which allow for some flexibility in the sense that inputs can be complementary or substitutes:

$$\theta_{i,t+1}^C = A[\gamma_1(\theta_{i,t}^C)^\rho + \gamma_2(\theta_i^N)^\rho + \gamma_3(I_{i,t})^\rho]^{\frac{1}{\rho}} e^{\eta_{i,t}} \tag{3.4}$$

$$A = exp(\alpha_0 + \alpha_1 X_i) \tag{3.5}$$

The total factor productivity (TFP) is represented by the parameter A, whereas the vector X_i accounts for the vector of observables of the child learning environment. $\theta_{i,t}^C$ represents the stock of cognitive skills in the previous period and θ_i^N accounts for non-cognitive skills. $I_{i,t}$ is the investment in early childhood, and $\eta_{i,t}$ represents the shocks affecting cognitive skill formation. The coefficients γ are the share parameters, such that we must have $\gamma_1 + \gamma_2 + \gamma_3 = 1$.

ρ is the parameter that indicates the degree of substitution among inputs in the skill production. Inputs can go from perfect complements ($\rho \to -\infty$) to perfect substitutes ($\rho \to 1$). If ρ approaches zero, the production function is a Cobb-Douglas. The elasticity substitution is given by $\sigma = \frac{1}{(1-\rho)}$. We assume the same technology for cognitive skill formation during all childhood.

3.2 Early childhood investment function

Investments in the child human capital reflect family choices, and might depend on parents' preferences, family resources, on the cost of investment, among other factors. We assume that parents have a desire to positively affect the child's future human capital, and choose the amount of investment subject to the household characteristics and the

[1] As of Cunha, Heckman and Schennach (2010), Attanasio et al. (2015),Attanasio, Meghir and Nix (2015) and Attanasio et al. (2017)

child's human capital. Although we do not derive a complete structural model [2], we briefly discuss intertemporal family choice in appendix F.

Parental investment is commonly treated in the literature as an unobserved latent factor, just like the skills (ATTANASIO et al., 2017; ATTANASIO; MEGHIR; NIX, 2015), that summarizes all the material and time investments devoted to the child. Given that our focus is to evaluate the effects of a specific investment - preschool attendance - we define investment as a variable that reflects parental decision on how many years of daycare and kindergarten to provide.

The number of preschool years attended by child i, PS_i, is the sum of the years spent in daycare (DC_i), from 0-3 years old, and kindergarten (KD_i), at ages 4-5:

$$PS_i = DC_i + KD_i \tag{3.6}$$

$$DC_i \in \{0, 1, 2, 3, 4\} \tag{3.7}$$

$$KD_i \in \{0, 1, 2\} \tag{3.8}$$

A child can experience at most 6 years of preschool, if she is enrolled in daycare during the first year of life, spending 4 years in daycare and 2 years in kindergarten. A child enrolled in daycare at the age 3 attends 3 years of preschool (a year in daycare plus two in kindergarten), being 3 years exclusively under home care. If the child is enrolled only in the last year of kindergarten, she had spent 5 years of home care and experienced a year of preschool. Finally, a child that started school only at mandatory age (6 years old) experienced 6 years of home care and no preschool. We assume that once enrolled in daycare or kindergarten, the child does not get back to home care and return to elementary school [3].

We consider investment as a linear combination of the number of years exclusively under home care (HC_i) and the number of years in daycare (DC_i) and kindergarten (KD_i):

$$I_{i,1} = \alpha_1 * HC_i + \alpha_2 * DC_i + \alpha_3 * KD_i \tag{3.9}$$

$$HC_i + DC_i + KD_i = 6 \text{ and } HC_i \in \{0, \ldots, 6\} \tag{3.10}$$

Parental choice on $I_{i,1}$ would surely depend on the characteristics of the household, such as family income, the availability of a person to take care of the child at home, the

[2] Some previous works in this literature (as Attanasio et al. (2017) and Attanasio et al. (2015)) do not explicitly model the family investment decision, and consider instead a reduced-form equation of parental investment.

[3] This is not a strong assumption for Brazilian data, as noted by Santos (2016) and Pinto, Santos and Guimarães (2017)

number of children in the family, and so on. In addition, the decision on investment would be affected by the current level of the child's human capital $h_{i,t} = g(\theta^C_{i,t}, \theta^N_i)$ [4]. Parents could decide to allocate more investment to more skilled children, reinforcing current status, or to provide more investment to the children with lower human capital, in a compensating fashion.

Parents would then choose the amount of investment in early childhood according to the following log-linear equation [5]

$$ln(I_{i,1}) = \lambda_0 + \lambda_1 Z_i + \lambda_2 ln(h_{i,1}) + v_i \tag{3.11}$$

where Z_i represent the household observables and $h_{i,1}$ the child human capital. Characteristics of the household that affect investment decision (Z_i) can differ from the learning context that affects cognitive skill production (X_i). Both X_i and Z_i are assumed to be fixed overtime.

Investment would be endogenous if parental decision is affected by omitted inputs of the production function. For instance, there can be a dimension of income we cannot control for that affects the child development and parental decision on investment. Moreover, parents can react to unobserved shocks affecting skill accumulation. If the child is hit by a negative shock that delays development, parents might compensate by providing more investment, or reinforce the negative shock by reducing investment. If this is the case, then $E(\eta_{i,t}|I_{i,t}) \neq 0$.

To explicitly account for endogeneity, we allow the shock of the production function to be correlated with the shock of the investment equation. More specificaly, we assume that conditional on all the observables, this correlationn is linear:

$$E(\eta_{i,t}|X_i, Z_i, \theta^N, \theta^C_{t-1}) = c\ v_i \tag{3.12}$$

To address endogeneity of investment in skill formation, we pose the existence of an instrument d_i, an observable that affects parents' decision but is uncorrelated to skill formation. If our instrument is valid, we would be able to isolate the exogenous part of investment decision. We could then think of the observables as including the instrument, being $Z_i = [z_i, d_i]$ Finally, we assume $v_i \sim N(0, \sigma_v)$ and v_i independent of Z_i in equation 3.11.

[4] We do not rule if parents decide on investment based in cognitive or non-cognitive skills, but allow flexibility to be any function of both skills. Moreover, we allow for the possibility that parents do not clearly identify skills θ^C and θ^N, but a function of them, imperfectly perceiving the child human capital at young ages.

[5] We model investment choice in a log-linear specification as of Cunha, Heckman and Schennach (2010) and Attanasio, Meghir and Nix (2015)

We then correct for endogeneity by running the investment equation as a first stage, computing the estimated residual $\hat{v}_i$ and including it as a regressor in the CES production function. The cognitive production function (equation 3.4), in logs, becomes

$$ln(\theta^C_{i,t+1}) = \frac{1}{\rho} ln[\gamma_1(\theta^C_{i,t})^\rho + \gamma_2(\theta^N_i)^\rho + \gamma_3(I_{I,t})^\rho] + \alpha_0 + \alpha_1 X_i + \Phi\hat{v}_i + \tilde{\eta}_{i,t} \tag{3.13}$$

where Φ is the coefficient of the control function. To sum up, the cognitive skill by the age when the child enters mandatory education ($\theta^C_{i,2}$) would be affected by the child's learning environment, early childhood investment and by the child's non-cognitive skills, as described above.

3.3 Measurement system

The latent ability $\theta_{i,t} = (\theta^C_{i,t}, \theta^N_i)$ is not directly observable and can only be measured with error. Ignoring measurement error would lead to several bias (ATTANASIO; MEGHIR; NIX, 2015; CUNHA; HECKMAN; SCHENNACH, 2010). Following Cunha and Heckman (2010), we consider a dedicated measurement system for skills, such that cognitive measures are affected by cognitive skills and non-cognitive measures by non-cognitive skills. Assuming a semi-log relationship among measures and latent factors, we have

$$M^C_t = \mu^C_t + \lambda^C_t ln(\theta^C_t) + \varepsilon^C_t \tag{3.14}$$

$$M^N = \mu^N + \lambda^N ln(\theta^N) + \varepsilon^N \tag{3.15}$$

where M^C_t is the vector representing cognitive skills' measures at a given period of time and M^N represents non-cognitive skills' measures (because we consider non-cognitive factor to be stable over childhood, we omit the time subscript). μ^C_t and μ^N are the vectors of intercepts, λ^C_t and λ^N the factor loadings. The measurement errors ε^C_t and ε^N are assumed to be independent of the latent abilities and normally distributed.

The number of measures required for the measurement system to be identified depend on the assumptions made about how the factors θ^C_t and θ^N interact in the skill production process (CARNEIRO; HANSEN; HECKMAN, 2003), and on the assumptions on the measurement errors (CUNHA; HECKMAN; SCHENNACH, 2010). We assume errors to be independent among measures in the same factor and across factors, and also serially uncorrelated. As of Attanasio, Meghir and Nix (2015), our system is identified if we have at least two measures per factor in each period, provided that at one factor has at least three measures [6].

[6] It is possible to relax some of these assumptions having more than two measures per factor in each period, as shown in Cunha, Heckman and Schennach (2010).

Our measurement system can be summarized as below, such that Λ is the matrix of factor loadings and Σ a diagonal matrix:

$$M = \Lambda ln(\theta) + \Sigma\varepsilon \tag{3.16}$$

Because abilities have no absolute measures [7], we normalize the factor loading of the first measure in each group to 1 (CARNEIRO; HANSEN; HECKMAN, 2003). For instance, if we have three measures of cognitive skills in a given period, $m^C_{1,t}, m^C_{2,t}$ and $m^C_{3,t}$, we set $\lambda^C_{1,t} = 1$, and the factor loadings $\lambda^C_{2,t}$ and $\lambda^C_{3,t}$ will be relative to the first measure.

In reality, we demean and scale all the measures for computational issues (as shown in section 5.2). For instance, $\tilde{m}^C_{j,t}$ are the demeaned cognitive measures

$$\frac{\tilde{m}^C_{j,t}}{\sigma^C_{j,t}} = \frac{\lambda^C_{j,t}}{\sigma^C_{j,t}}\ ln(\theta^C_t) + \frac{\varepsilon^C_t}{\sigma^C_{j,t}} \tag{3.17}$$

and we actually normalize $\lambda^{*C}_{1,t} = \frac{\lambda^C_{1,t}}{\sigma^C_{1,t}} = 1$. This is the same of normalizing the factor loadings of the first (original) measure to $\frac{1}{\sigma^C_{1,t}}$ instead of 1. The same goes for the non-cognitive measures.

[7] Cunha, Heckman and Schennach (2010) avoid the problem of not having an intuitive and feasible unit of measure for skills by anchoring all measurements' factor loadings in terms of adult wages. We do not have available data to do so.

4 Data

We use a longitudinal panel dataset due to Felicio et al. (2013), with data collected in 2008 and 2012 for the city of Sertãozinho, in São Paulo state. Sertãozinho's population is about 110,000 inhabitants[1], being also a relatively rich city: the R$ 39,477 GDP per capita (current prices, year of 2015) is higher than national average, putting the city among the 10% highest municipal GDPs in Brazil, and top 20% in the state of São Paulo. Formal workers' average wages are also above national average, placing Sertãozinho in position 154 comparing to the other 5,570 Brazilian municipalities [2].

Sertãozinho's education system also performs well when compared to other municipalities. On a scale from 0-10 in IDEB (Basic Education Development Index), an index that comprehends Math and Portuguese proficiency exams and the flow of students that are not retained in grade, Sertãozinho got a 6.7 for students in the first years of elementary school. This represents position 341 out of 5,570 municipalities, and position 97 among 645 municipalities in the state of São Paulo [3].

In 2008, Felicio et al. (2013) applied a literacy exam (Provinha Brasil) to every children attending the 2^{nd} year of elementary education in public and private schools in Sertãozinho [4]. This exam evaluates child's literacy, such that the maximum score would be reached if the child is fully able to read and write. While children answered Provinha Brasil, their parents (or the adult responsible for the child) answered a questionnaire informing demographic and socioeconomic characterization of the household, in addition to questions about daycare and kindergarten enrollment. We have in our database 2,076 children that took Provinha Brasil exam and had parent's questionnaire filled.

In the year of 2012 an effort was made to follow the same children that had participated in 2008, in order to collect further information from them. Four years later, they should be enrolled in the 6^{th}, but some children that were retained were attending 5^{th} or 4^{th} instead. In the first half of 2012 (2012.1), Math and Portuguese proficiency tests were applied to every student enrolled in the 6^{th} and 5^{th} grade, and to students in every classroom of 4^{th} grade that had at least one child that had participated in 2008. Proficiency test scores were computed for 3,359 students, since this round included many students that

[1] 110,074 according to IBGE latest Census of 2010, and an estimated population of 122,643 for the year of 2017 - https://cidades.ibge.gov.br/brasil/sp/sertaozinho/panorama

[2] https://cidades.ibge.gov.br/brasil/sp/sertaozinho/panorama

[3] Source: INEP, http://ideb.inep.gov.br/

[4] Provinha Brasil is usually applied to students in public schools only, and participation is not mandatory (as seen in http://portal.inep.gov.br/provinha-brasil). The partnership with the Secretary of Education and INEP made it possible to apply to every student that was in school in the date Provinha Brasil was applied

had not participated in the 2008 survey. The tests were designed under IRT (Item Response Theory), such that standardized scores are comparable across grades. Parents (or other adult responsible for the child) answered an extensive take home questionnaire providing detailed information about the context of the child and the household, socioeconomic status, among other information. 2,749 children had this questionnaire returned by parents.

In the second half of 2012 (2012.2), all 4,522 students enrolled in $6^{th}, 5^{th}$ and 4^{th} grade answered a questionnaire with four blocks of items, that intended to assess socioemotional skills. Instruments from the psychology literature were adapted to Brazilian context and to children age (about 12-13 years old), such as the Big Five Personality Traits (Openness, Conscientiousness, Extroversion, Agreeableness, Neurocitism), developed by Caprara et al. (1993).

The sample considered in our analysis consists of the 1,044 children that took Provinha Brasil (in 2008), Math and Portuguese proficiency tests (in 2012.1), answered non-cognitive questionnaire (in 2012.2), and had parent's questionnaires filled in both 2008 and 2012. The main reasons for a child that participated in 2008 not be found in 2012 are absence in school in the day the exams were taken, inconsistency in administrative data (the main source to identify the children among questionnaires was child's name, such that having conflicting names would make it not possible to match all the questionnaires), or even moving to another city. If none of those sources is correlated to our investment variable (years of daycare/kindergarten attendance), our analysis would still be valid.

Santos (2016) points to an estimated attrition between 2008 and 2012 of about 20%. Fonseca (2015) checked for the presence of selection effects, testing if the survival rate between 2008 and 2012 was affected by initial conditions and preschool enrollment. In a simple OLS exercise, she regressed the literacy scores in 2008 against dummy variables for preschool attendance and for survival in 2012, to test whether children that survived in the database until 2012 or attended preschool already had higher Literacy Scores in 2008. Although it was not possible to reject these null hypothesis, the dummy that interacted both effects was not significant, indicating that attrition among preschool attendees and non-attendees would not have been different. Although Fonseca (2015) was interested in a different process, tracking differences between 2008 and 2012 test scores, the point here is that there seems to be no evidence of systematic attrition bias due to selection effects. The reasons for students that participated in 2008 had not participated in 2012 are arguably not related to parental investment in the form of preschool attendance.

In our case, we use information from all questionnaires to build the investment variable and the instrument. Moreover, we do not directly consider Provinha (2008) and proficiency scores (2012) as abilities, but use raw data to generate a synthetic database and draw the distribution of latent factors, which are the ones used in our models. Therefore, given our context, there are not further meaningful tests for selection bias we can perform

in raw data.

In the second semester of 2017, another field survey was conducted, to gather further information of the same students in Sertãozinho. Once more, there were applied socioeconomic questionnaires, Math and Portuguese proficiency exams and a questionnaire to assess socioemotional skills. Preliminar information of students of our 2008-2012 sample that were mapped in 2017 are used to run an extra exercise: we consider Portuguese and Math scores of 2017 as a measure for cognitive skills in the long-term. We do not fully discuss the presence of attrition bias in 2008-2012-2017 panel since the data collected in 2017 is still being processed.

4.1 Descriptive Statistics

Table 1 presents mean and standard deviation of some variables of our dataset. In average, children start school at daycare or kindergarten have more educated parents, with a higher proportion of parents that have completed high school or entered college, meanwhile most parents from children with no preschool did not finish high school [5].

Children only enrolled in school at mandatory age also come from lower income households, with a higher proportion of families receiving social benefits, and lower socioeconomic index than full sample average.

Families that have more children as less likely to enroll their child in preschool, either. This is in line with the quantity-quality tradeoff, so that investment is costly and households with a larger number of children would invest less in each child.

Most of the average differences among children that did not attend any preschool to children enrolled in daycare or kindergarten are statistically significant, as shown in appendix tables 35 and 36.

4.2 Measures for latent skills

We consider three measures for cognitive skills in the period that ranges from 2008 to 2012, considered in our model as a single period: Literacy (Provinha Brasil) scores and Portuguese and Math proficiency scores.

Provinha Brasil consists in two exams of reading and writing, such that the score is the simple average of the number of correct answers out of 24 questions in each exam. The scale ranges from 0 to 24, being the maximum score achieved if the child is fully literate. Due to the methodology and objectives, the resulting distribution is asymmetric and clearly not normal, since many children got the maximum score. It is also possible to

[5] Parents that did not attend school and missings are omitted from the table

Table 1: Household and child characteristics, by school entry age

	Full Sample		**Daycare**		**Kindergarten**		**Elementary**	
Observations	1044		364		550		130	
Descriptive Statistics	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age	7.59	0.61	7.55	0.57	7.59	0.61	7.72	0.67
Male	0.49	0.50	0.45	0.50	0.51	0.50	0.51	0.50
White	0.43	0.49	0.44	0.50	0.44	0.50	0.34	0.48
Mother Education (highest level achieved)								
Elementary School	0.39	0.49	0.33	0.47	0.42	0.49	0.43	0.50
Middle School	0.21	0.41	0.20	0.40	0.23	0.42	0.16	0.37
High school	0.23	0.42	0.28	0.45	0.23	0.42	0.08	0.27
College	0.04	0.20	0.08	0.27	0.03	0.17	0.00	0.00
Father Education (highest level achieved)								
Elementary School	0.36	0.48	0.32	0.47	0.38	0.49	0.42	0.50
Middle School	0.22	0.41	0.21	0.41	0.23	0.42	0.17	0.37
High school	0.23	0.42	0.27	0.45	0.23	0.42	0.13	0.33
College	0.04	0.21	0.09	0.28	0.03	0.16	0.00	0.00
Household configuration								
Lives with both parents	0.66	0.47	0.64	0.48	0.70	0.46	0.59	0.49
Number of siblings	2.09	2.00	1.95	1.96	2.00	1.89	2.90	2.37
Has a dictionary at home	0.69	0.46	0.69	0.46	0.71	0.45	0.57	0.50
A shelf with 20+ books	0.27	0.45	0.29	0.45	0.29	0.45	0.17	0.38
Has children's books	0.76	0.43	0.79	0.41	0.77	0.42	0.65	0.48
Attended public school in 2008	0.91	0.29	0.86	0.35	0.92	0.27	0.98	0.15
Dimensions of income								
Bolsa Familia beneficiary	0.22	0.41	0.18	0.39	0.20	0.40	0.39	0.49
Other social benefit	0.09	0.28	0.08	0.27	0.07	0.26	0.15	0.36
SES Index (calculated from data)	11.60	4.72	12.46	5.23	11.68	4.37	8.90	3.74

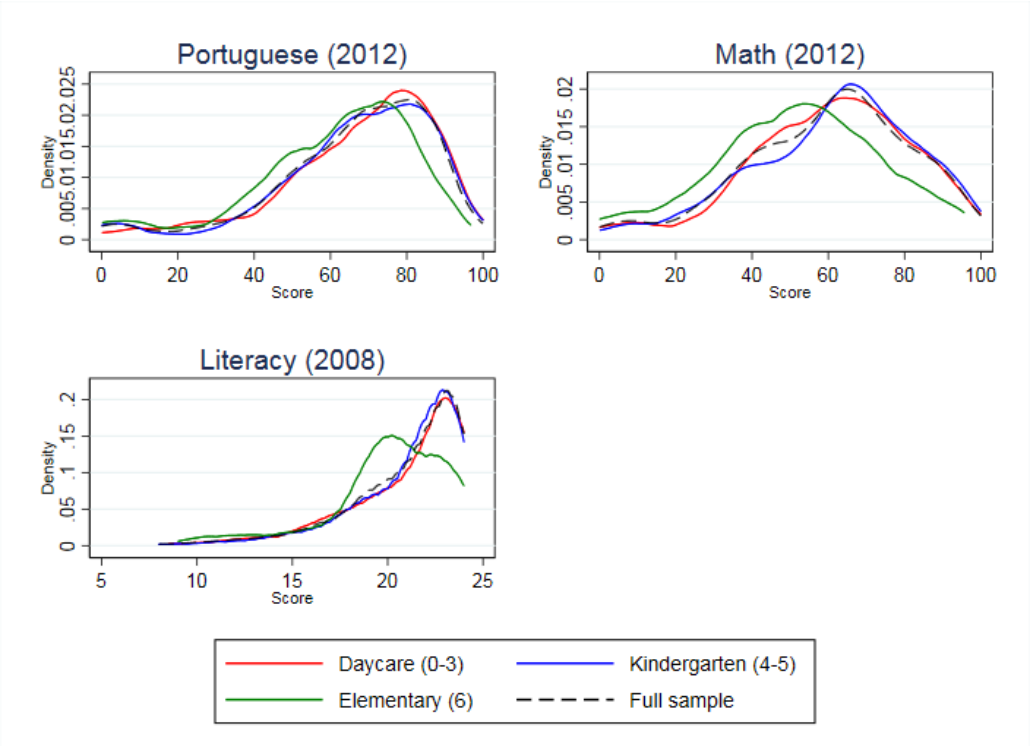


Figure 1: Portuguese and Math Proficiency Scores (2012) and Literacy Scores (Provinha 2008) distributions, by school entry age

transform Provinha's scores in a scale compatible with SAEB proficiency that ranges from 217.41 to 665.04 points (FONSECA, 2015).

Portuguese and Math tests were designed under IRT, and the scores are already standardized in a scale from 0 to 100 points.

Literacy and Portuguese and Math exams have different purposes and are measured in different scales, being not fully comparable [6]. Since we estimate a latent factor using factor analysis, it is not such a problem to use measures in different scales [7]. We also report all of our results using only 2012 proficiency scores as measures for cognitive skills (without Literacy) in appendix B.1.

It is clear from figure 1 that students that were only enrolled in school in elementary education at 6 years old (green line) performed poorly when compared to daycare and kindergarten attendees in all available measures of cognitive skills. Indeed, the mean scores of daycare/kindergarten attendees are significantly different from non-attendees, and we do reject the null that the distribution of their scores is the same, using Kolmogorov test (as presented in the appendix tables 35 and 36.

As measures of non-cognitive skills we use the constructs of the BFI [8]. The five constructs are obtained from a list of 44 statements, each designed to assess one of the five personality traits. The five construct were computed from their respective statements using Cronbach's Alpha (CRONBACH, 1951).

The picture that emerges from socioemotional measures is much less clear. Children that had been under homecare only would have lower Extraversion and Oppeness. However, most of the differences across distributions are not significant, as shown in Kolmogorov tests reported in the appendix D tables. Although we do not rule out the possibility that preschool attendance could affect non-cognitive skills, we do not model it.

Because the BFI constructs are built from a self-reported questionnaire, we should be aware of the possibility of acquiescence bias, which is the tendency of agreeing with a given statement, even when in doubt [9]. Figure 3 contains the full sample distribution of the raw constructs extracted from the questionnaire and adjusted for acquiescence bias. Although we opt for the raw BFI in our baseline estimation, we report the results using

[6] A discussion on this topic is found in Fonseca (2015)

[7] Other works using this approach put together measures reflecting different aspects of child, even continuous test scales and dummies, such as Attanasio, Meghir and Nix (2015),Attanasio et al. (2017) and Attanasio et al. (2015).

[8] DeYoung (2010) and DeYoung et al. (2010) work provide evidence on the biological sources, as of areas of the brain related to the Big Five Personality Traits.

[9] If a person is consistent in the degree of agreeableness, in a scale from 1 to 5, being 1 for strongly disagree and 5 for strongly agree with a given statement, we should observe an average of 3 for statements that have exactly the opposite direction on the same subject. The within-subject adjustment is performed by computing mean of opposite statements for each individual, and subtracting it from each answer. This way all answers adjusted at the individual level would be in a comparable scale (FISCHER, 2004).

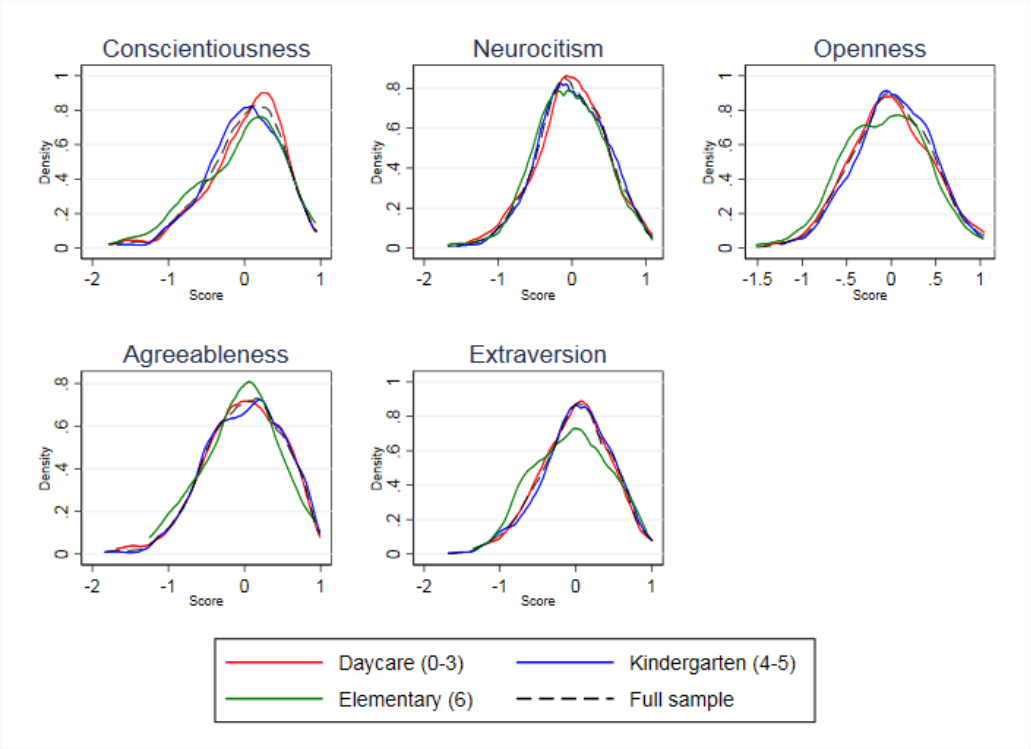


Figure 2: Big Five Personality Traits distribution, by school entry age

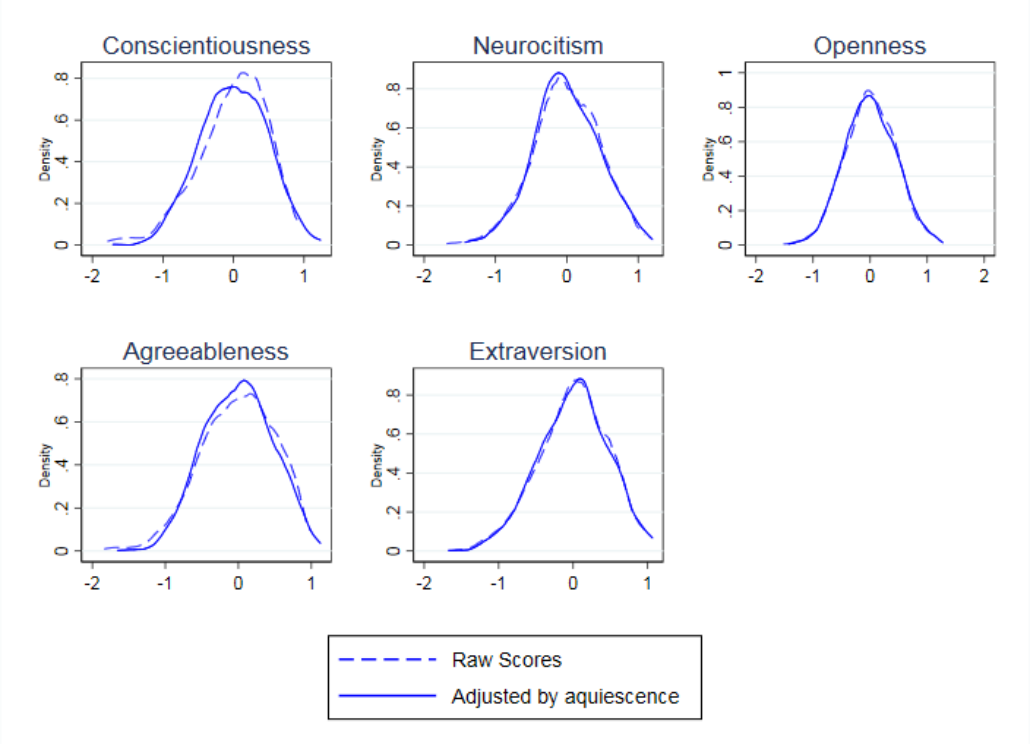


Figure 3: Big Five Personality Traits corrected by acquiescence at the individual level

measures for the non-cognitive skill corrected for acquiescence in appendix B.2.

We also test measuring socioemotional skills by another group of measures, the Social Skills: Cooperation, Assertiveness, Empathy and Self-control. Interestingly, results are pretty similar to the obtained using the BFI constructs (see appendix B.3).

4.3 Instrument

It is quite difficult to find a suitable instrument for preschool attendance in Brazil. Pinto, Santos and Guimarães (2017) use the daycare supply variation at a municipal level as instrument. In our case, we need a variable that influence parent's decision on enrolling the child in daycare and kindergarten that does not affect ability formation through any other channel and varies at individual level, since the entire data is from a single municipality[10].

We instrument the investment by the distance to the closest Early Childhood Center (ECC), which is considered as a *proxy* of the availability of daycare/kindergarten nearby the household that the child lived [11]. The proximity of an ECC would positively influence parents to enroll their child in daycare/kindergarten, and families that have to cover longer distances to an ECC are less likely to enroll their child in preschool [12].

To calculate the distance of the household to the nearest ECC, we first had to estimated the area where the child lived (due to compliance issues, it was not possible to collect the exact address of the family in the field survey). We use the location of the elementary school each children attended in the 2^{nd} year of mandatory education along with the amount of time (in minutes) taken to reach elementary school and main transportation mode used (both reported by parents). Combining these information with the average speed reached in each transportation mode collected from Google API, we are able to estimate the distance (in meters) that each child lived from elementary school. We use this distance as the radius where the household where the child lived would have lived.

[10] We tried using as instrument two dummy variables from our database. The first indicated if any grandmother lived at a distance of at most 20 minutes from the household, and the second indicated if anyone in the household was a beneficiary from Bolsa Família Program, what would give priority in public daycare's waiting list. Both candidates to instrument presented low correlation to our investment variable, plus low variation at individual level. We were not able to use them in this empirical framework, because the EM algorithm would not converge in various specifications tested, so results are not reported.

[11] It is a *proxy* for availability since it does not reflect precisely the number of vacancies in the nearest ECC by the time the child was aged 0-6 years old. For the case of public ECCs, when there is no room available, the family can put their child in the waiting list by signing a form and bringing a copy of the child's documents. This process is specific for each ECC, and there is no unified waiting list or record of the number of children in the waiting list for the ECCs, such that the exact shortage of spots in the nearest ECC cannot be recovered.

[12] Barros et al. (2011) found that distance was one of the main drivers for daycare enrollment decision (for data in Rio de Janeiro, Brazil), and that parents could not properly distinguish daycare centers according to their quality.

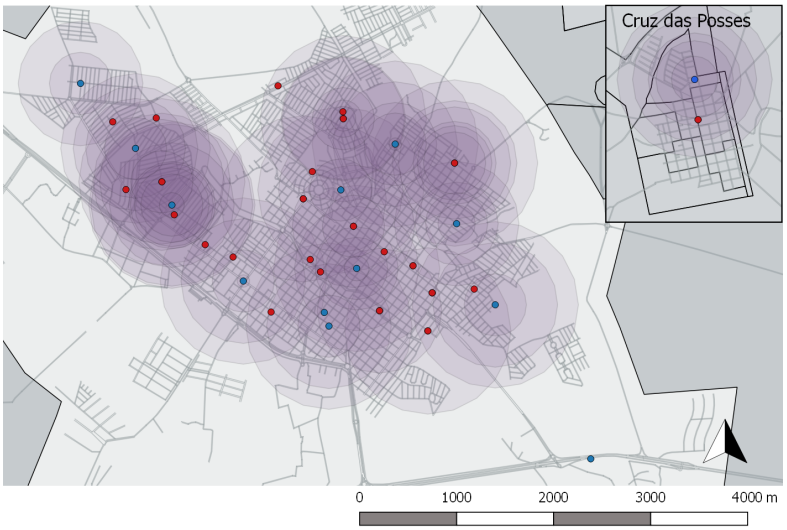


Figure 4: Elementary schools attended (blue), estimated radius where students lived and ECC (red dots)

Importantly, we assume that the family did not move between early childhood and the 2^{nd} year of elementary school.

Figure 4 displays the location of elementary schools that the 1044 children in our sample attended in 2008 (blue dots) and the greatest area in which children that attended each elementary school would have lived. For ease of viewing, only areas with radius up to 1.1 km are displayed. 25.5% of the children in our sample lived at a distance of at most 1km from elementary school, and nearly half of the sample at most 1.5 km away from the school attended in 2008, as shown in appendix E tables.

Figure 5 plots the 27 public and private centers offering preschool in the city of Sertãozinho in the year of 2008 [13]. ECCs are relatively well distributed across the city. In the centre of the figure we have the main part of the city, being the light grey the urban area. On the right and on the top Cruz das Posses district, which is 12 km away from the center of Sertãozinho.

Our instrument is the distance from the border of the radius to the closest ECC (red dots). The appendix E describes in further detail every step taken and all the calculations.

[13] The children in the sample could have attended preschool between the years of 2000 and 2007. We did not include schools that were opened after 2008.

Figure 5: Early Childhood Centers in Sertãozinho-SP - source: IBGE

4.4 Control Variables

We include as controls the variables that capture aspects that characterize the learning environment of the child and are assumed to be free of measurement error [14].

The empirical framework used demands parsimonious specification regarding the number of controls. Otherwise, the EM algorithm used to generate the synthetic dataset in the first step does not converge. Therefore we select a few variables form our rich dataset to summarize child's learning environment.

Socioeconomic Status Index (SES): computed from parent's answers from the socioeconomic questionnaire (answered in 2008) regarding household characteristics and appliances, in a scale adapted from Critério Brasil, also for the year of 2008 [15].

Parents' schooling: we consider the highest schooling among the mother and father (or woman/man responsible for the child, respectively) reported in 2008 questionnaire. Since we do not have the exact number of school years attended, we consider a step variable corresponding to the highest education level fully completed, being: 0 years if did not attend school neither completed elementary school; 4 if completed elementary school

[14] Our empirical framework does require to assume that all the remaining variables, except for the measures of the latent factors, are measured without error. Our control variables rely entirely on self-reported questionnaires (filled by parents and students) and could be subject to some kind of error. However, most these answers refer to noncontroversial matters, and there does not seem to be any conflict of interest of respondents or some potential source of systematic bias in this answers.

[15] A number of points is attributed according to the facilities of the household (such as the number of fridges, washing machines, DVD players, automobiles, the presence of a housekeeper) and the education level of the household principal. The index is a discrete punctuation covering all of the items, and is shown to be highly correlated to Brazilian income distribution. Source: http://www.abep.org/criterio-brasil

but not middle school; 8 for middle school and high school dropouts; 11 for the ones that finished high school and 15 for complete college.

Gender and Race: dummy variables that equal 1 if male and if white, respectively. Both items come from the questionnaire answered by students in 2012 [16].

Number of siblings: number of brothers and sisters the child had in 2008, when in the 2^{nd} year of elementary education.

Age: we included the age the child had in 2008, obtained from cross-checking across 2008 and 2012 questionnaires.

Parent's marital status: a dummy variable whether the parents had been married during all years that correspond to when the child was aged 0 to 6. This variable is obtained from 2012 parent's questionnaire.

Serious illness during childhood: a dummy variable that equals 1 if the child has been seriously ill in any year during early childhood, from 2012 parent's questionnaire.

[16] Parents also answered about race in both 2008 and 2012 socioeconomic questionnaires, but due to small differences in the answers we opted to consider how the students consider themselves regarding to race.

5 Empirical Strategy

5.1 Model adapted to the available data

We estimate a simpler version of our model with the available data, considering three periods in childhood. Period $t = 1$ comprehends early childhood, when investment takes place. In period $t = 2$ we observe the measures for the latent cognitive and non-cognitive skills. In period $t = 3$ the long-term outcome is observed.

In period $t = 1$ parents decide on early childhood investment, I_i. Parents take into account child's human capital endowment $h_{i,0} = \theta_{i,0}$ and the household's characteristics Z_i.

The existing literature for Brazilian data usually considers daycare and kindergarten enrollment as a dummy variable treatment, comparing preschool attendees with the appropriate controls. We consider preschool as an investment, measured as a continuous variable, in an attempt to capture the marginal effect of the time spent in preschool (not only the binary attendance), thus putting attendees and non-attendees in the same scale.

In our dataset we can distinguish whether the child was enrolled in daycare at the age 3, but we do not know exactly at which age children enrolled from 0-2 years old started attending daycare. The maximum number of preschool years is therefore truncated in 4, such that DC_{0-2}, DC_3 are dummies:

$$PS_i = DC_{0-2i} + DC_{3i} + KD_i \tag{5.1}$$

$$DC_{0-2i} \in \{0, 1\} \tag{5.2}$$

$$DC_{3i} \in \{0, 1\} \tag{5.3}$$

$$KD_i \in \{0, 1, 2\} \tag{5.4}$$

We pose the investment variable to be measured as follows:

$$I_{i,1} = \alpha_1 * HC_i + \alpha_2 * DC_{0-2i} + \alpha_3 * DC_{3i} + \alpha_4 * KD_i \tag{5.5}$$

$$HC_i + DC_{0-2i} + DC_{3i} + KD_i = 6 \text{ and } HC_i \in \{0, \ldots, 6\} \tag{5.6}$$

In order to run a log-linear equation for parental investment decision of how many preschool years to provide, we set the $\alpha_1 = \frac{1}{6}$ in equation 5.5 such that when the child attended no preschool ($HC_i = 6$), we have $ln(I_i) = 0$. We calibrate the remaining values relative to α_1 , being $\alpha_2 = \alpha_3 = (1 + \alpha_1)$ and $\alpha_4 = (1 - \alpha_1)$:

$$I_{i,1} = \frac{1}{6}HC_i + \left(1 - \frac{1}{6}\right)DC_{0-2_i} + \left(1 + \frac{1}{6}\right)(DC_{3i} + KD_i) \tag{5.7}$$

Graphically, $ln(I_{i,1})$ behaves as displayed in figure 6. The minimum value of the log of the investment is zero, if the child has remained all early childhood under home care. As long as the family trades years of full home care for some preschool, the investment rises. Even though we cannot split daycare years for ages 0-2 in our data, if we could observe the last two columns in the right, the log of the investment variable would still be rising for additional daycare years.

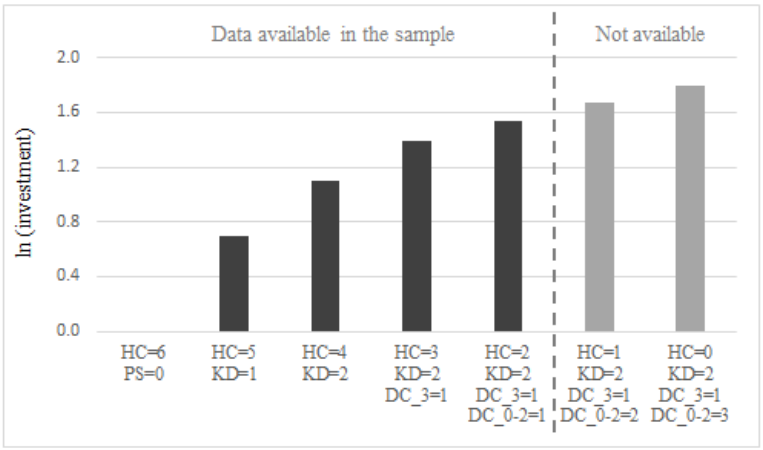


Figure 6: Investment variable (in log), according preschool attendance

Since the values of the α's were assigned *adhoc*, we analyze the sensibility of the estimates due to the investment variable specification by reporting in the appendix A results for different values of α_1 to α_4.

The investment equation remains the same, except that we have no information on θ_0 in the equation below, then term $\lambda_2 ln(\theta_{i,0})$ is contained in the estimated residual. The investment equation is given below, and we assume $v \sim N(0, \sigma_v)$:

$$lnI_{i,1} = \lambda_0 + \lambda_1 Z_i + \lambda_2 ln(\theta_{i,0}) + v_i \tag{5.8}$$

With available data, we are able to estimate the production function of cognitive skills considering substitution among two inputs: non-cognitive skills and early childhood investment:

$$\theta_{i,2}^C = exp(\alpha_0 + \alpha_1 X_i)[\gamma {I_{i,1}}^\rho + (1-\gamma)\theta_i^{N\rho}]^{\frac{1}{\rho}} e^{\eta_{i,1}} \tag{5.9}$$

Investment choice might be endogenous in the skill production function if parents' choice is correlated to omitted inputs or to unobserved shocks affecting child development. We address endogeneity by using as instrument the distance to the nearest child care center offering preschool or kindergarten to the (estimated) household location.

We run the investment equation (5.8) with the instrument and include its residual in the cognitive production function:

$$ln(\theta_{i,2}^C) = \frac{1}{\rho} ln[\gamma {I_{i,1}}^{\rho} + (1-\gamma)\theta_i^{N\rho}] + (\alpha_0 + \alpha_1 X_i) + \Phi\hat{v}_i + \tilde{\eta}_{i,1} \tag{5.10}$$

In the following period, no investment decision is made. The production of cognitive skill throughout the late childhood and adolescence years (period $t = 3$) considers substitution among the previous cognitive skill level ($\theta_{i,2}^C$) and socioemotional skill (θ_i^N). There is no direct effect of preschool attendance $I_{i,1}$ other than through its impact in $\theta_{i,2}^C$.

There is a substantial body in the literature pointing that the timing of investment in childhood matters (CUNHA; HECKMAN, 2007; CUNHA; HECKMAN, 2010; HECKMAN, 2006) and that the productivity of inputs might differ between early and later stages (HECKMAN; STIXRUD; URZUA, 2006). Henceforth, we allow the productivity (the coefficients β), the share coefficients (δ) and the substitution parameter (ϱ) to be different between early childhood and later childhood.

Cognitive skill for period $t = 3$ is given by:

$$\theta_{i,3}^C = exp(\beta_0 + \beta_1 X_i)\ [\delta {\theta_{i,2}^C}^{\varrho} + (1-\delta)\theta_i^{N\varrho}]^{\frac{1}{\varrho}} e^{\eta_{i,2}} \tag{5.11}$$

5.2 Estimation Strategy

We follow the estimation method proposed by Attanasio, Meghir and Nix (2015), that builds on Cunha, Heckman and Schennach (2010). We assume that the joint distribution of the log of the latent factors is a mixture of two normals, say A and B, being τ the mixing parameter and $\Phi(\mu, \Omega)$ the CDF of the normal distribution:

$$F_\theta = \tau\Phi(\mu_A, \Omega_A) + (1-\tau)\Phi(\mu_B, \Omega_B) \tag{5.12}$$

Departing from a mixture of normals is crucial to allow our skill production function to be a CES. The production function can be seen as the conditional expectation of the inputs. If the log of the factors followed a normal distribution, which has linear conditional

mean, we would restrict our production function to be linear or Cobb-Douglas (which is linear in logs). It is possible to consider a mixture of three or more normal distributions, to allow for greater flexibility, which we intend to do in future work.

We do not observe the latent factors directly, but only their corresponding measures. Given the structure of our measurement system (equations 3.14 and 3.15), the distribution of the measures will also be a mixture of log-normals [1]:

$$F_M = \tau\Phi(\Pi_A, \Psi_A) + (1-\tau)\Phi(\Pi_B, \Psi_B) \tag{5.13}$$

In fact, we augment the distribution to include the set of controls, which are assumed to have no measurement error. We include the controls X by setting the $\Sigma\varepsilon = 0$ and the corresponding factor loadings to 1. This step is necessary to account for all relevant variances and covariances among measures and controls in our original data.

$$F_{M,X} = \tau\Phi(\Pi_A^{M,X}, \Psi_A^{M,X}) + (1-\tau)\Phi(\Pi_B^{M,X}, \Psi_B^{M,X}) \tag{5.14}$$

In the first step, the parameters of the distribution $F_{M,X}$ are recovered from the data using Maximum Likelihood Estimation (MLE). We use the Expectation Maximization (EM) Algorithm [2] to identify the parameters of $F_{M,X}$ from the data [3]. First, given our sample, we use k-means clustering to guess initial values for the parameters of the distribution $F_{M,X}$, and based on them, the "E" step of the algorithm estimates the probability that each observation belongs to distribution A or B, from which we compute an estimate for the mixture τ. Then, in the "M" step we update the likelihood function and get new estimates for $\Pi_A^{M,X}, \Pi_B^{M,X}, \Psi_A^{M,X}, \Psi_B^{M,X}$, from which the "E" step generates another guess for τ and so on. The algorithm iterates "E"and "M" steps until the relative growth of the likelihood function is smaller than the convergence criteria [4]. In order to speed up convergence, before running the first step we normalize the measures for the latent factors and all the controls by subtracting the mean and dividing by the standard deviation.

In the second step, we use minimum distance to recover the parameters of the joint distribution of the factors. We choose the factor loadings (matrix Λ) [5] so as to minimize

[1] $M = \Lambda ln\theta + \Sigma\varepsilon$. Then $\Phi_k = \Lambda'\Omega_k\Lambda + \Sigma$ and $\Pi = \Lambda\mu_k$, $k = \{A, B\}$.

[2] From Dempster, Laird and Rubin (1977) and improvements of Arcidiacono and Jones (2003).

[3] In principle, we could use MLE to and try to jointly identify the parameters τ and $\Pi_A^{M,X}, \Pi_B^{M,X}, \Psi_A^{M,X}, \Psi_B^{M,X}$ from our sample, which is an independent draw of $F_{M,X}$. However, this procedure is computationally demanding and we use the EM algorithm to speed up convergence

[4] Further explanation of how the iteration of EM algorithm converges to the maximum of the log likelihood function is found in Train (2009).

[5] Recalling that the factor loading of the first measure in cognitive and non-cognitive groups and loading of all controls are set to 1.

the distance between the means and variance-covariance matrices of the distribution of the measures (5.14) and the distribution of the latent factors (5.15).

$$F_{\theta,X} = \tau\Phi(\mu_A^{\theta,X}, \Omega_A^{\theta,X}) + (1-\tau)\Phi(\mu_B^{\theta,X}, \Omega_B^{\theta,X}) \tag{5.15}$$

In order to run investment and skill production functions, we generate simulated data from $F_{\theta,X}$, which is the distribution of interest. We build a synthetic dataset with 10,000 draws from this distribution, that is considered as data to run the regressions for our analysis. The synthetic dataset is supposed to contain all relevant information provided by the original sample (the 1044 observations longitudinal panel data).

The investment function is estimated using OLS, and residuals are stored to be used as a control function in the CES cognitive skill production. The parameters of the CES function are estimated using Non-inear Least Squares (NLS). We report results with and without the control function from the investment equation.

6 Results

First we estimate the joint distribution of measures and controls. From our original 1044 observations sample we generate a synthetic dataset with 10,000 observations, assuming the data generation process described in section 5.2. We estimate the log of the latent cognitive and non-cognitive factors, $ln(\theta^C)$ and $ln(\theta^N)$ (recovered from our measurement system), together with the log of the investment, $ln(I)$ and the controls, accounting for the variances and covariances presented in the original data [1]. The mixing parameter τ for our baseline specification is estimated to be 0.89, with a 95% confidence interval of $[0.09; 0.91]$, obtained using 500 bootstrap samples with replacement.

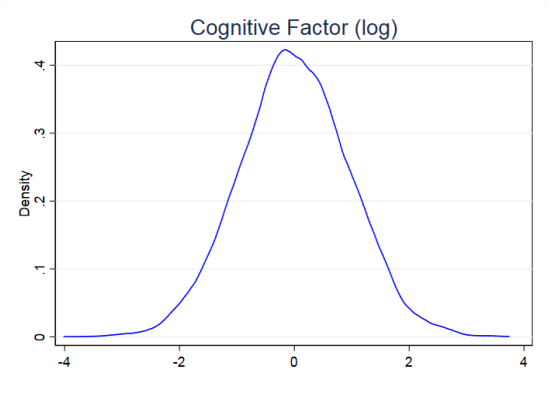
The distributions of the latent factors drawn from our synthetic dataset are plotted in figure 7. We assume that the measures are related to the log of the latent factors (equations 3.14 and 3.15), and normalize the cognitive and non-cognitive factors in units of Portuguese Score and of the Conscientiousness construct, respectively. The distribution of these measures are displayed along with the distributions of the latent factors, as a reference.

The factor loadings are reported in table 2. Among the measures for cognition, Portuguese score presents the highest factor loading in all specifications tested. Literacy scores (from 2008) are not as informative about the cognitive factor as Math and Portuguese scores (from 2012). Even though excluding Literacy would not significantly affect our estimates (as discussed in appendix B.1), we prefer to include Literacy (Provinha) scores to account for all available information in our original data.

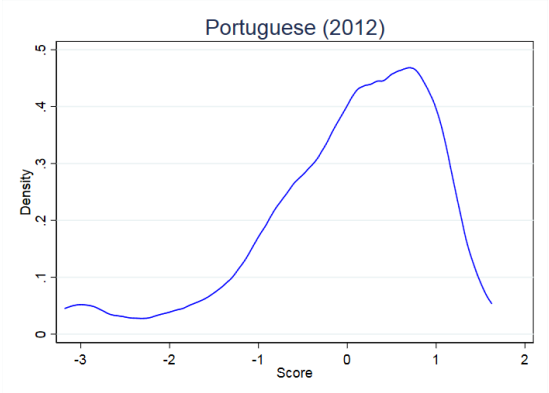
The constructs of the BFI seem to be equally informative about the non-cognitive factor, being Neurocitism and Extraversion slightly below the others. We opted to use the raw constructs of the BFI, but also run the entire estimation considering measures corrected for acquiescence bias, and report results in appendix B.2. Finally, we consider other group of four measures for the socioemotional factor and draw a very similar latent factor from them, and get also similar results for the CES equation (further detail in appendix B.3).

We also decompose the variance of the measures between the part that is related to the variance of the latent factor, and the variance of the measurement error. From the measurement system, the variance of a measure $m^k_{j^k}$ in the vector of measures $M^k,\ k \in \{C, N\}$ and is given by

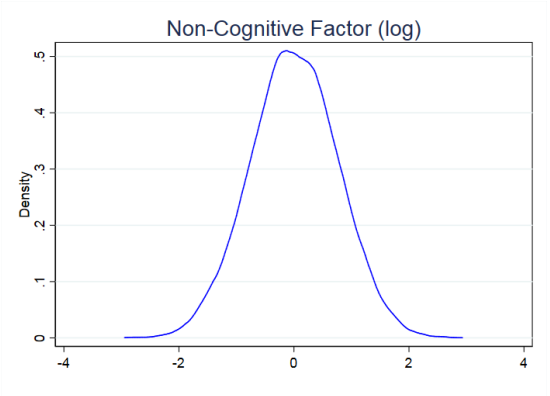
[1] Even the observations with from the original dataset that had missing values in some of the variables of interest were considered to generate the synthetic dataset, that is from now on considered as data.



(a) Latent factors from synthetic dataset (10,000 obs)



(b) Normalized measures (1044 obs)



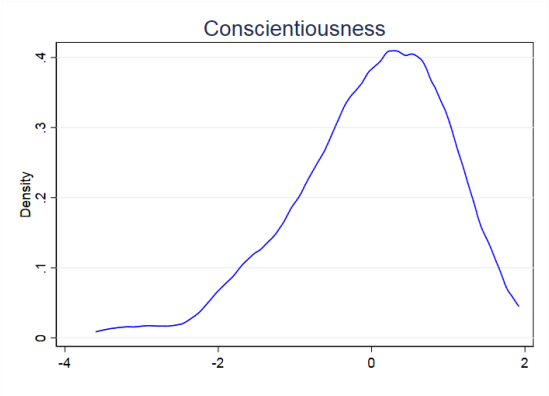


Figure 7: Distributions of latent factors and normalized measures

Table 2: Factor Loadings

Latent Factor	Measures	Loading	Signal to Noise
Cognitive skills	Portuguese Score	1.00	89.5%
	Math Score	0.86	68.7%
	Literacy Score	0.57	31.1%
Non-cognitive skills	Conscientiousness	1.00	60.6%
	Neurocitism	0.90	49.4%
	Openness	1.00	59.0%
	Agreeableness	1.04	62.6%
	Extraversion	0.91	48.9%

$$var(m^k_{j^k}) = (\lambda^k_{j^k})^2 var(ln\theta^k) + var\varepsilon^k_{j^k} \tag{6.1}$$

$$s^{\theta^k}_{j^k} = \frac{(\lambda^k_{j^k})^2 var(ln\theta^k)}{(\lambda^k_{j^k})^2 var(ln\theta^k) + var\varepsilon^k_{j^k}} \tag{6.2}$$

The above ratio is denominated signal to noise ratio [2], and gives a sensibility of how informative our measures are regarding the latent factors we are interested in. Indeed, our measures are informative, with signals close to or above 50%. According to the methodology used, a substantial part of the variance of the measures for abilities is due to measurement error, since $s^{\varepsilon}_{j^k} = 1 - s^{\theta^k}_{j^k}$.

6.1 Investment Equation

Using OLS, we regress the log of the investment in early childhood against the instrument and controls. We report investment coefficients of our baseline specification with and without age as a control variable - columns (1) and (2), respectively.

Age is a variable that we would like to control for in the skill production function, but its role in the investment equation is not clear. Since the controls of the CES production function have to be a subset of the controls of the investment equation (controls that affect investment decision but not the CES operate as instruments), we run both investment and CES functions with and without age as a control. Our baseline specification includes age in both equations, but estimates do not change dramatically if we exclude it - coefficients of columns (1) and (2) are stable.

Our guess for age to have a negative impact on investment is that in our raw data, children that started school in elementary school only were older than the average of the sample. This might be due to the fact that the enforcement of the change of school entry age from 7 to 6 years (when elementary school duration was augmented from 8 to 9 years) was not uniform in Sertãozinho in the year of 2008. Then we could have children that did attend preschool from 0-6 years and were enrolled at school only at the age of 7, and then a negative correlation among age and investment.

The instrument has a negative signal, as expected, meaning that increasing the distance to the nearest Early Childhood Center (ECC) would lead to lower investment during

[2] Recall that we assume $ln\theta \sim \tau N(\mu^A, \Omega^A) + (1-\tau)N(\mu^B, \Omega^B)$, being the corresponding variance $var(ln\theta) = \tau\Omega^A + (1-\tau)\Omega^B + \tau(1-\tau)(\mu^A - \mu^B)(\mu^A - \mu^B)^T$. Our estimate of signal to noise differs from the code provided by Attanasio et al. (2017), that does not include the term $\tau(1-\tau)(\mu^A - \mu^B)(\mu^A - \mu^B)^T$ when calculating the variance of $ln\theta$, which we believe is missing. This term is not set to zero by imposing the zero mean normalization, i.e., making $\tau\mu^A + (1-\tau)\mu^B = 0$, as posed by Attanasio, Meghir and Nix (2015).

Table 3: Investment Equation Coefficients

Investment equation coefficients	(1) Baseline	(2) Baseline, no age
Instrument (smallest distance to ECC)	-0.048	-0.048
	[-0.1 ; 0.01]	[-0.1 ; 0.01]
Socioeconomic Index (SES)	0.116	0.112
	[0.005 ; 0.18]	[0.004 ; 0.17]
Male	-0.028	-0.032
	[-0.08 ; 0.04]	[-0.08 ; 0.03]
White	0.023	0.024
	[-0.05 ; 0.07]	[-0.05 ; 0.07]
Age	-0.052	
	[-0.1 ; 0.03]	
Number of siblings	-0.096	-0.101
	[-0.14 ; 0]	[-0.15 ; -0.01]
Parents' highest education	0.148	0.157
	[0.09 ; 0.26]	[0.1 ; 0.26]
Parents married in early childhood	-0.047	-0.045
	[-0.11 ; 0.03]	[-0.11 ; 0.03]
Seriously ill in early childhood	-0.019	-0.017
	[-0.08 ; 0.04]	[-0.08 ; 0.05]

Note: columns (1) and (2) correspond to the same synthetic dataset. 95% confidence intervals obtained from 500 bootstrap replications in square brackets.

preschool years. Even with all limitations imposed by the data (as discussed in appendix E), the instrument we built is found to be negative and significant in all specifications tested in this work. However, even if the distance to the closest daycare/kindergarten center is an exogenous shock, it might not be strong enough to address the endogeneity of parental decision, that has many dimensions. This issue is further explored in section 6.3, where we report results excluding the instrument from all the estimation.

Income could have an ambiguous effect over preschool attendance: higher income parents could care more about investing in their child, but also afford to spend more time with the child at home or provide another arrangement - such as hiring a baby sitter and delaying school entry. In our data we find that children raised in a household with a higher *proxy* for socioeconomic status have higher chances of attending some preschool. Attanasio et al. (2017) and Attanasio, Meghir and Nix (2015) also estimated a positive effect of income over investment factors, although considering different dimensions of investment as ours.

From the aspect of the household configuration, married couples would also be able to afford delaying school entry if one parent can stay longer with the child at home Irrespective of socioeconomic and marital status, higher educated parents would invest more in providing preschool years than less educated ones, and this is the highest coefficient in all specifications for investment equations. Furthermore, according to our

model, increasing the number of siblings would lower the investment level, which is plausible with the quantity-quality tradeoff, and also found by Attanasio, Meghir and Nix (2015) and Attanasio et al. (2015).

Finally, we tried to capture another dimension affecting parental decision which is child health status. Parents could be afraid that enrolling a child at school before mandatory age could cause her to be ill due to the contact with other children and adults at the daycare/kindergarten. Moreover, a child that had been seriously ill might be not even be able to attend school. Indeed, we find that a child that had serious illness in any year from age 0 to 5 would attend less preschool years. However, health is likely to be an endogenous variable, highly dependent of parental inputs, income, access to health care, or even parent's skills (ATTANASIO et al., 2017; ATTANASIO; MEGHIR; NIX, 2015). In short, health status could be an output of parental investment. Despite of that, being *seriously* injured could have an exogenous component.

6.2 CES Production Function

We estimate the CES production function using NLS. Since we have no measure of cognitive skills prior to early childhood investment, we are not able to recover the previous latent factor of cognitive skills ($\theta_{i,0}$). This way we cannot control for previous skill nor investigate self-productivity or persistence in the production of cognitive skills in period $t = 2$.

We do find evidence of cross-productivity, with non-cognitive skills affecting the production of cognitive skills, and of complementarity between investment and non-cognitive skills in the cognitive skills formation.

The CES function estimates are displayed in table 4. Columns (1) and (2) are estimated in the same synthetic database, while the third pair of columns is run in a different synthetic dataset, without marital status and the presence of serious illness as control variables. If we do not include the control functions (columns b), the share of the investment is estimated at about 30% in the production of cognitive skills.

The estimated complementarity parameter (ρ) is close to zero and all the confidence intervals contain zero, such that we cannot reject that the production functions are of Cobb-Douglas form - a result that is in line with findings in this literature. Although our estimates could be inaccurate, due to limitations from the data and from the empirical framework already discussed, we find a not so high substitutability among inputs, with an elasticity of substitution close to unity. This is an evidence that early childhood investment matters for cognitive skill formation [3] and cannot be fully compensated by non-cognitive

[3] A result also found in the literature, such as Attanasio, Meghir and Nix (2015), Attanasio et al. (2015),Heckman, Stixrud and Urzua (2006)

Table 4: CES Production Function Coefficients

CES coefficients	(1) Baseline		(2) Baseline, no age	
	(a) CF	(b) no CF	(a) CF	(b) no CF
Investment (γ)	0.715	0.289	0.718	0.292
	[0.64;0.84]	[0.24;0.37]	[0.65;0.84]	[0.24;0.38]
Non-cognitive skills ($1-\gamma$)	0.285	0.711	0.282	0.708
	[0.16;0.36]	[0.63;0.76]	[0.16;0.35]	[0.62;0.76]
Complementarity (ρ)	0.013	0.009	0.015	0.012
	[-0.29;0.09]	[-0.14;0.08]	[-0.29;0.09]	[-0.14;0.09]
Elasticity Substitution (σ)	1.013	1.009	1.015	1.012
	[0.77;1.1]	[0.88;1.09]	[0.77;1.1]	[0.88;1.09]
Control Function (CF)	-0.668		-0.671	
	[-0.82;-0.59]		[-0.82;-0.59]	
Intercept	-0.001	-0.002	-0.001	-0.002
	[-0.02;0.04]	[-0.02;0.03]	[-0.02;0.04]	[-0.02;0.03]
Socioeconomic Index (SES)	-0.069	-0.015	-0.071	-0.019
	[-0.16;0.05]	[-0.11;0.08]	[-0.17;0.05]	[-0.11;0.07]
Male	-0.088	-0.026	-0.091	-0.031
	[-0.16;-0.02]	[-0.09;0.05]	[-0.16;-0.02]	[-0.1;0.04]
White	0.065	0.035	0.066	0.037
	[0.01;0.15]	[-0.03;0.11]	[0.01;0.15]	[-0.03;0.12]
Age	-0.028	-0.055		
	[-0.13;0.04]	[-0.15;0]		
Number of siblings	-0.045	-0.060	-0.047	-0.064
	[-0.13;0.02]	[-0.14;0.02]	[-0.14;0.02]	[-0.14;0.01]
Parents' highest education	0.085	0.134	0.090	0.144
	[-0.04;0.17]	[0.03;0.21]	[-0.04;0.18]	[0.04;0.23]

Note: all estimates for baseline specification run in the same synthetic dataset. The CF of (1.a) and (2.a) are the residuals from the respective investment equations (with and without age as a control). 95% confidence intervals obtained using 500 bootstrap replications in square brackets.

skills.

In column (1a) the control function (CF) is the residual of the investment equation that includes age, and in (2a) the residual of investment equation (2), that does not account for age. The CF are negative and significant in all specifications tested, and causes the estimated share of investment to raise to more than 70%.

Negative signs for the CF were also found by Attanasio et al. (2015) for very young children, and by Attanasio, Meghir and Nix (2015) and Attanasio et al. (2017) for children at different ages living in diverse villages and regions. These authors attribute the negative impact as parents responding to negative shocks that affect the child development by providing more investment. This is valid if we account for all relevant variables that influence parents' decision and the instrument fully addresses the endogeneity, such that the residual in the investment equation contains only exogenous shocks.

This is not an assumption we could rely on for our data. Moreover, we are dealing with children in the same municipality, that has a relatively homogeneous education system.

It is harder to argue that we could have those negative shocks affecting children in our sample in individual level. If we are not able to fully solve for endogeneity, omitted inputs remain in the residual of the investment equation.

We believe the negative control function indicates that parents choose to compensate lower initial skills by providing more investment to the child. Furthermore, a mechanism that we do not account for is whether parents consider preschool enrollment as a substitute or complement for their own time spent with the children.

Parents' education has a significant positive effect both in cognitive skill formation and in the investment decision. However, in the CES equation zero is contained in all confidence intervals from this variable, indicating that the effect of parental education over child's cognitive formation is more likely to occur by affecting the investment decision than through a direct effect itself. These findings are consistent with previous works for Brazilian data, where parent's education is found to be positively correlated with the probability of attending preschool and with proficiency scores (PINTO; SANTOS; GUIMARÃES, 2017; CURI; MENEZES-FILHO et al., 2006).

Surprisingly, income is found to have a negligible and negative effect over cognitive formation (with all confidence intervals containing zero), what goes against existing evidence in the literature. Our results suggest that income affects investment decision, but not directly cognitive formation.

The number of siblings of each child could capture not only the negative aspect of diluting parent's limited time and resources, but also the *stimuli* of one child on another. In our model, the net effect over cognitive skill formation is estimated to be negative.

From all of discussed, we consider the estimates obtained without the control function (columns b) to be more reasonable and stable than columns (a). We believe our instrument we cannot fully solve for the endogeneity of investment decision and, when including the residual of the investment equation as a control function, we are overestimating the impact of investment over cognition.

6.3 Further discussion on the instrument

Our instrument could be endogenous to preschool attendance if parents concerned with child development had chosen to live in neighborhoods with greater supply of daycare and kindergarten. Although we do not have means to control for this issue with available data, we argue that the 27 schools offering daycare and/or kindergarten are reasonably spread across the city of Sertãozinho, being few areas unattended.

However, we do not control for is preschool quality. Our instrument would be endogenous if parents choose household location in order to be closer to better preschools.

If the instrument is endogenous to preschool attendance, we are likely to have an upward bias.

Our exclusion restriction would fail if the distance to the closest ECC directly affected cognition development in any other way than through investment. The household location is not randomly assigned, depending on family resources and preferences, and that many features of the neighborhood can somehow affect child development. With available information we cannot assert that poorer and disadvantadge areas have systematically less ECCs than more favorable areas or the other way around. In this sense, the distance to the closest center offering preschool is could be orthogonal to cognitive skill production.

Table 5: Investment Equation Coefficients - baseline and specifications without instrument

Investment equation coefficients	(1) Baseline	(3) No instr.	(3) No instr, no age	(4) No instr, fewer controls	(4) No instr, fewer controls, no age
Instrument (smallest distance to ECC)	-0.048 [-0.1 ; 0.01]				
Socioeconomic Index (SES)	0.116 [0 ; 0.18]	0.091 [-0.01 ; 0.17]	0.088 [-0.01 ; 0.17]	0.063 [0 ; 0.16]	0.060 [0 ; 0.15]
Male	-0.028 [-0.08 ; 0.04]	-0.016 [-0.08 ; 0.03]	-0.022 [-0.09 ; 0.03]	-0.019 [-0.08 ; 0.04]	-0.021 [-0.08 ; 0.03]
White	0.023 [-0.05 ; 0.07]	-0.013 [-0.06 ; 0.06]	-0.012 [-0.06 ; 0.06]	0.002 [-0.06 ; 0.06]	0.003 [-0.06 ; 0.06]
Age	-0.052 [-0.1 ; 0.03]	-0.059 [-0.1 ; 0.02]		-0.041 [-0.11 ; 0.02]	
Number of siblings	-0.096 [-0.14 ; 0]	-0.090 [-0.14 ; -0.01]	-0.095 [-0.15 ; -0.01]	-0.083 [-0.14 ; 0.01]	-0.087 [-0.15 ; 0]
Parents' highest education	0.148 [0.09 ; 0.26]	0.168 [0.08 ; 0.27]	0.179 [0.09 ; 0.28]	0.188 [0.09 ; 0.27]	0.196 [0.1 ; 0.27]
Parents married in early childhood	-0.047 [-0.11 ; 0.03]	-0.034 [-0.11 ; 0.02]	-0.033 [-0.11 ; 0.03]		
Seriously ill in early childhood	-0.019 [-0.08 ; 0.04]	-0.011 [-0.08 ; 0.05]	-0.008 [-0.08 ; 0.05]		

Note: columns (1), (3) and (4) correspond to different synthetic datasets. 95% confidence intervals obtained from 500 bootstrap replications in square brackets.

In table 5 we compare the baseline specification to the coefficients obtained when we run the whole estimation without the instrument. In columns (3) we exclude the main instrument but still have parent's marital status and child health in the early childhood working as instruments (because included in the investment equation and not in the CES). In columns (4), we exclude all the variables that are working as instruments, and still have stable coefficients comparing to the baseline, both with and without including age as a control. From this exercise we can only conclude that our instrument's covariance with the other controls is low.

For the CES equation, we compare the baseline without control function (1.b) to the equations run in synthetic datasets that do not include the instrument. Even setting aside the other variables that work as instuments as well (columns 4), we still get stable

Table 6: CES Production Function Coefficients - baseline and specifications without instrument

CES coefficients	(1) Baseline		Without instrument, no CF			
	(a) CF	(b) no CF	(3) No instr.	(3) No instr, no age	(4) No instr, fewer controls	(4) No instr, fewer controls, no age
Investment (γ)	0.715 [0.64;0.84]	0.289 [0.24;0.37]	0.311 [0.23;0.38]	0.314 [0.23;0.38]	0.351 [0.28;0.4]	0.354 [0.28;0.41]
Non-cognitive skills (1-γ)	0.285 [0.16;0.36]	0.711 [0.63;0.76]	0.689 [0.62;0.77]	0.686 [0.62;0.77]	0.649 [0.6;0.72]	0.646 [0.59;0.72]
Complementarity (ρ)	0.013 [-0.29;0.09]	0.009 [-0.14;0.08]	0.054 [-0.13;0.1]	0.057 [-0.13;0.09]	-0.171 [-0.31;0.06]	-0.170 [-0.32;0.06]
Elasticity Substitution (σ)	1.013 [0.77;1.1]	1.009 [0.88;1.09]	1.057 [0.88;1.11]	1.060 [0.88;1.1]	0.854 [0.76;1.07]	0.855 [0.76;1.06]
Control Function (CF)	-0.668 [-0.82;-0.59]					
Intercept	-0.001 [-0.02;0.04]	-0.002 [-0.02;0.03]	0.003 [-0.02;0.03]	0.002 [-0.02;0.03]	0.037 [-0.01;0.06]	0.037 [-0.01;0.06]
Socioeconomic Index (SES)	-0.069 [-0.16;0.05]	-0.015 [-0.11;0.08]	-0.032 [-0.1;0.09]	-0.036 [-0.1;0.08]	0.003 [-0.11;0.09]	-0.001 [-0.11;0.08]
Male	-0.088 [-0.16;-0.02]	-0.026 [-0.09;0.05]	-0.021 [-0.09;0.05]	-0.028 [-0.09;0.04]	-0.027 [-0.11;0.04]	-0.031 [-0.11;0.03]
White	0.065 [0.01;0.15]	0.035 [-0.03;0.11]	0.050 [-0.03;0.11]	0.052 [-0.03;0.11]	0.028 [-0.03;0.11]	0.030 [-0.03;0.11]
Age	-0.028 [-0.13;0.04]	-0.055 [-0.15;0]	-0.066 [-0.14;-0.01]		-0.069 [-0.14;0]	
Number of siblings	-0.045 [-0.13;0.02]	-0.060 [-0.14;0.02]	-0.051 [-0.14;0.02]	-0.057 [-0.15;0.01]	-0.072 [-0.17;0.01]	-0.079 [-0.17;-0.01]
Parents' highest education	0.085 [-0.04;0.17]	0.134 [0.03;0.21]	0.137 [0.03;0.2]	0.149 [0.04;0.22]	0.140 [0.04;0.22]	0.153 [0.05;0.23]

Note: columns (1), (3) and (4) correspond to different synthetic datasets. 95% confidence intervals obtained from 500 bootstrap replications in square brackets.

coefficients for the CES, and conclude that our excluded instruments do not directly affect cognitive skill formation.

Furthermore, our instrument could be exogenous, but not strong enough to move parental decision on preschool attendance. In a city with distances no greater than 30 minutes by car, the distance to schools offering daycare and kindergarten might not play a important role in affecting the probability of going to preschool.

6.4 Long-term

We use proficiency collected in 2017 as measures for long-term outcome. In the data that is already available from the survey, we were able to identify 706 out of the 1044 students that participated in 2008 and 2012 [4]. There were applied different Portuguese

[4] This number should be seen as a lower bound and is likely to grow as long as collected information is fully processed. 919 students from the 2008-2012 panel were identified in the schools of Sertãozinho prior to the field survey was conducted, using schools' administrative data and School Census. Yet there is a possibility of attrition (i.e., students with lower ability could have dropped out school), so far the main reason for a student have not been identified is unmatched names in different forms,

and Math exams, whether children were attending elementary or high school years.

We report descriptive statistics of this subgroup relative to our full sample in appendix D.1. The 706 students found had performed above the entire 1044 observations sample in Portuguese and Math exams in 2012, but with no clear distinction with respect to socioemotional measures. Yet, we estimate the synthetic database using only these 706 students as raw data (instead of the full sample) and run the investment and CES equations. We still get similar coefficients as of the baseline model, as shown in appendix B.4. Although we are not able to run estimtions including the long-term cognition for the entire sample, this exercise indicate that the results obtained from this 706 observations is likely to be representative for the entire sample.

As measures for cognition in 2017 we use the raw number of questions answered correctly in Portuguese and Math exams, whose distribution is displayed in figure 8. We consider that all questions have the same weight, since the scores have not been adjusted by IRT yet. This implies, for instance, that we are assuming that a correct answer in the Math exam applied to elementary school students is equivalent to a correct answer in the high school Math exam for a children attending high school [5].

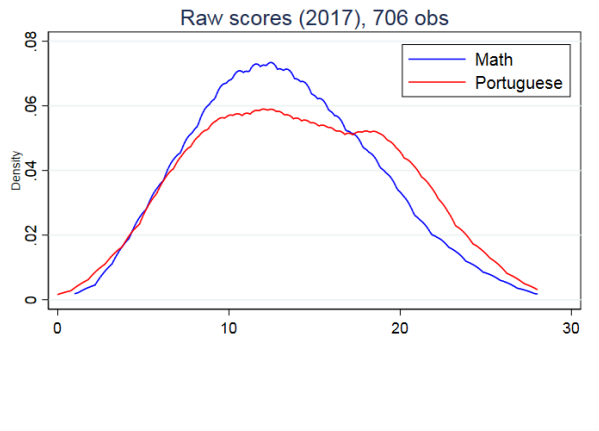


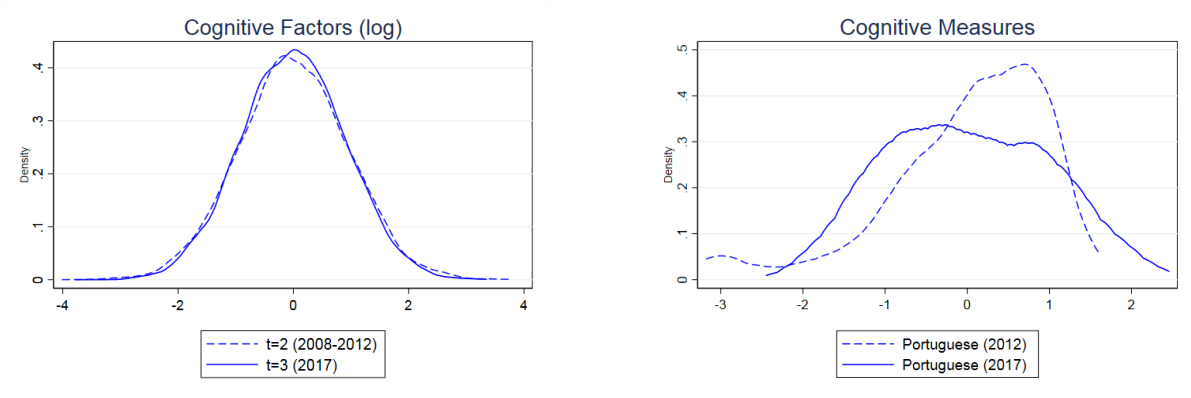
Figure 8: Distributions of Portuguese and Math exams applied in 2017

We estimate the synthetic dataset adding the measures for θ_3^C together with the measures for the latent factors θ_2^C and θ^N, and the same set of controls of the baseline specification. The mixture parameter τ is estimated to be 0.10, with a 95% confidence interval of $[0.08; 0.91]$.

The factor loadings are reported in table 7. Portuguese proficiency remains as the measure with the highest factor loading among measures for cognition.

issues with administrative data or even the presence of homonyms

[5] Nearly half of our sample already had experienced at least one in-grade retention between 2008 and 2012, and there were students that accumulated retentions between 2012 and 2017. If a school year contributes to cognitive skill formation, we could be underestimating the cognition of retained students.



(a) Latent factors from synthetic dataset (10,000 obs)

(b) Corresponding normalized measures

Figure 9: Distributions of the short (θ_2^C) and long-term (θ_3^C) latent factors, and of the normalized measures

Table 7: Factor Loadings - including long-term cognitive measures

Latent Factor	Measures	Loading	Signal to Noise
Cognitive skills (t=2)	Portuguese Score	1.00	77.9%
	Math Score	0.93	70.9%
	Literacy Score	0.65	38.3%
Cognitive skills (t=3)	Portuguese Score	1.00	76.9%
	Math Score	0.59	31.7%
Non-cognitive skills	Conscientiousness	1.00	55.1%
	Neurocitism	0.92	48.3%
	Openness	1.08	62.8%
	Agreeableness	1.05	60.3%
	Extraversion	0.93	46.9%

For the sake of brevity we omit the estimates of the investment equation , since the parameters are barely the same from baseline specification.

In the table 8 we present estimates for the CES production functions of θ_2^C (with and without Control Function), which remain similar to our baseline specification (that is, departing from full 1044 observations sample, without including measures for long-term cognition). The coefficients of the production of θ_3^C are reported in the third column and should be taken with caution, given the caveats of the data already discussed.

We find strong evidence of self-productivity, with previous cognition (θ_2^C) accounting for a 82.1% share in the production of θ_3^C. We might not fully capture the importance of non-cognitive skills since the factor considered was drawn from measures collected when children were aged about 12, that might not reflect the socioemotional skills in adolescence years. Assuming that the socioemotional factor remained stable during childhood is more plausible than assuming stability from childhood to teenage years.

Table 8: Short and Long-term CES production functions

CES coefficients: short and long-term	θ_C^2	θ_C^2, no CF	θ_C^3
Investment (γ)	0.793	0.293	
	[0.64 ; 0.89]	[0.2 ; 0.39]	
Non-cognitive skills ($1-\gamma$)	0.207	0.707	
	[0.11 ; 0.36]	[0.61 ; 0.8]	
Previous cognitive skill (δ)			0.821
			[0.72 ; 0.97]
Non-cognitive skills ($1-\delta$)			0.179
			[0.03 ; 0.28]
Complementarity (ρ)	0.081	0.048	-0.069
	[-0.7 ; 0.13]	[-0.32 ; 0.12]	[-0.17 ; 2.02]
Elasticity Substitution (σ)	1.088	1.051	0.935
	[0.58 ; 1.15]	[0.76 ; 1.13]	[-11.54 ; 11.49]
Control Function (CF)	-0.781		
	[-0.89 ; -0.59]		
Intercept	-0.010	-0.006	0.008
	[-0.02 ; 0.09]	[-0.03 ; 0.06]	[-0.09 ; 0.01]
Socioeconomic Index (SES)	-0.060	0.012	0.022
	[-0.16 ; 0.07]	[-0.1 ; 0.13]	[-0.07 ; 0.14]
Male	-0.022	0.041	0.057
	[-0.1 ; 0.09]	[-0.04 ; 0.14]	[-0.02 ; 0.13]
White	0.040	0.001	0.017
	[-0.05 ; 0.14]	[-0.09 ; 0.09]	[-0.06 ; 0.09]
Age	-0.049	-0.106	-0.050
	[-0.15 ; 0.05]	[-0.18 ; 0]	[-0.12 ; 0.02]
Number of siblings	-0.032	-0.066	-0.018
	[-0.16 ; 0.06]	[-0.16 ; 0.04]	[-0.08 ; 0.07]
Parents' highest education	0.017	0.080	-0.025
	[-0.1 ; 0.15]	[-0.02 ; 0.2]	[-0.13 ; 0.09]

Note: 95% confidence intervals obtained using 500 bootstrap replications in square brackets

With available data, our model indicates that cognition in period $t = 2$ is determinant for cognition measured in adolescence years ($t = 3$). The estimated elasticity of substitution σ close to one keeps pointing to a Cobb-Douglas, but nevertheless with wide open confidence intervals.

The income has a positive effect over the production of cognition from childhood to adolescence years. The subset of students that were at school and took exams in 2017 are likely to have higher income than students that dropped or that students that work, for instance.

Overall changes in the magnitude and signal of the control variables from period $t = 2$ to $t = 3$ might also indicate that there could be other aspects of the learning environment affecting skill production that we are not accounting for.

The bottom line is that we find evidence of self-productivity in the formation of cognitive skills in teenage years. Although preschool effects for this sample can fade out overtime (FONSECA, 2015), we find that they are important due its impact in earlier cognition that is carried to later stages, thus having an indirect effect over the long-term

cognition.

An idea of the magnitude of this impact is given by the estimated coefficients in our model: 82.1% of cognition in period $t = 3$ is given by cognition in period $t = 2$, from which 29.3% would be due to investment (in the equation without the control function). If we multiply those shares, we get a compound effect of 24.0% for investment.

We can also estimate the contribution of investment to long-term cognition from the production functions. θ_3^C can be decomposed into the part that is due to inputs θ_3^C and θ_3^N, and θ_2^C into the effects of I_1 and θ^N:

$$\theta_3^C = \frac{\partial \theta_3^C}{\partial \theta_2^C}\theta_2^C + \frac{\partial \theta_3^C}{\partial \theta^N}\theta^N \tag{6.3}$$

$$\theta_2^C = \frac{\partial \theta_2^C}{\partial I_1} I_1 + \frac{\partial \theta_2^C}{\partial \theta^N}\theta^N \tag{6.4}$$

Combining both equations, we get [6]

$$\theta_3^C = \frac{\partial \theta_3^C}{\partial \theta_2^C}\frac{\partial \theta_2^C}{\partial I_1} I_1 + \left(\frac{\partial \theta_3^C}{\partial \theta_2^C}\frac{\partial \theta_2^C}{\partial \theta^N} + \frac{\partial \theta_3^C}{\partial \theta^N}\right)\theta^N \tag{6.5}$$

$$share_{I_1} = \frac{\partial \theta_3^C}{\partial \theta_2^C}\frac{\partial \theta_2^C}{\partial I_1}\frac{I_1}{\theta_3^C} \tag{6.6}$$

To compute $share_{I_1}$, we use the average values of the variables in the synthetic dataset. We use the distribution of the investment variable in the original dataset (706 observations), and compute the average values of for the same percentiles of the distribution of the investment variable in the synthetic dataset (along with the corresponding averages of the latent factors).

Table 9: Averages of synthetic dataset variables according to investment variable distribution

Original database				Synthetic database (average values for corresponding distribution of investment)			
Daycare (years)	Kindergarten (years)	Investment variable	Cumulative distribution	I_1	θ_2^C	θ^N	θ_3^C
0.0	0.0	1.0	10.48	0.19	1.26	1.35	1.26
0.0	1.0	2.0	29.60	0.44	1.45	1.46	1.45
0.0	2.0	3.0	64.16	0.96	1.47	1.52	1.47
1.0	2.0	4.0	90.37	2.19	1.57	1.57	1.57
2.0	3.0	4.7	100.00	6.62	1.74	1.56	1.74

Together with the parameters of the CES production functions, we compute the impact of investment in long-term cognition for each level of investment (table 10. When

[6] Full calculations in the appendix C

we use the coefficients of the CES with the control function (CF) for period $t = 2$, early childhood investment would explain about 63-65% of the level of the long-term cognition, and about 23-25% if we use the coefficients without the CF.

Table 10: Estimated effect of investment on long-term cognition

Original database			Share of investment		CAGR	
Daycare (years)	Kindergarten (years)	Investment variable	coef. with CF	coef. no CF	coef. with CF	coef. no CF
0.0	1.0	2.0	63.82%	23.20%	4.59%	1.91%
0.0	2.0	3.0	64.68%	23.81%	4.64%	1.96%
1.0	2.0	4.0	65.51%	24.46%	4.69%	2.01%
2.0	3.0	4.7	66.56%	25.33%	4.75%	2.07%

Although results are not directly comparable, Felício and Fernandes (2005) found an impact of preschool attendance over Math proficiency scores for students in the 4^{th} grade to be an average of 6.6% for SAEB 2003, using propensity score matching, and of an average of 9.6% for national data, using fixed effects to correct for the endogeneity of preschool. Being the estimated impact in proficiency measured 4 years after preschool, it would correspond to compound annual growth rates (CAGR) of 1.6% and 2.3% a year, if we consider a linear impact overtime.

In our data, long-term proficiency is measured 11 years after the preschool investment, that could have happened at most up to 2007 (when children in our sample would be already in the 1^{st} year of elementary school). This would lead to CAGR about 4.6-4.8% and 1.9-2.1%, whether we consider estimates with an without CF. Despite of the limitations of this comparison and the fact that the effects of preschool attendance are not likely to be linear, it seems reasonable to find higher impacts in a sample from a single municipality with better education standards than national average. This is consistent with findings in the literature that children from more favorable backgrounds and attend better quality schools benefit higher of preschool (PINTO; SANTOS; GUIMARÃES, 2017; SANTOS, 2016; VERAMENDI; URZÚA, 2011).

Again, these calculations must be seen as a rough approximation of an average effect to the available sample. In future work, we can use these framework to run a number of simulations and counterfactual exercises. For instance, we could simulate what would be the aggregated effect if every child in our sample had attended preschool, or whether preschool investment contributes to increase or reduce inequality among children from different backgrounds.

7 Final Remarks

We estimate the impacts of preschool attendance in cognitive skills using longitudinal panel data for children in Sertãozinho, SP. With a linear measurement system, we draw cognitive and non-cognitive unobserved latent factors and estimate CES cognitive skill production functions. We find that early investment greatly matters for the production of cognition during childhood, being complementary to non-cognitive skills. Cognition in the long-term (teenage years) is estimated to be highly persistent with respect to previous cognition, such that early childhood investment would indirectly affect the child development path in the following periods, which is consistent with the preceding literature. In line with previous works, we also find skill production functions to be Cobb-Douglas.

We consider preschool attendance as a continuous investment variable, that works as an input for the production of cognition, and is allowed to be endogenous. We instrument parental choice on investment by the distance to the nearest school offering daycare and/or kindergarten, and include the estimated residual of the investment equation as a control in the cognitive skill production function. Our instrument might be positively correlated to non observed dimensions of preschool quality and family income, thus the estimated impacts of preschool attendance over cognition can be biased upwards. A limitation of our model is considering early childhood (0-6 years old) as a single period. Parents' decision on daycare and kindergarten enrollment might have different sources of endogeneity that we are not able to capture.

An issue we do not address is the role of grade retention. Nearly half of the students in the longitudinal panel were retained between the years of 2008 and 2012. We do not consider retention as a measure of the latent factors, and use information of all students to generate the synthetic dataset in which we run the skill production functions, irrespective of the information that some students had been retained.

For future work, we believe that including measures for the quality of preschool could improve the instrument variable and the estimates of the investment equation. Moreover, when information from the field survey of 2017 is entirely available, it will be possible to use socioemotional measures collected in 2017 in long-term cognitive skill production, which we believe would improve our estimates. It will be also possible to include an equation of non-cognitive skill formation, that can provide further insights.

Finally, the estimated impacts of preschool attendance are above previous studies for Brazilian data, which is probably related due the the fact that we use data for a single municipality, with higher income and better education system than Brazilian average. As of Pinto, Santos and Guimarães (2017), a higher effect of preschool attendance could be

correlated with dimensions of quality of the preschool we are not controlling for. Having said that, our estimates should be taken as an upper bound of the impact of early childhood investment. Even with all the limitations discussed above, the reported estimates are stable and consistent across all specifications tested, and the message is quite clear: preschool attendance matters for cognitive skill formation in the childhood, and since cognition is persistent across stages, investment has an indirect cumulative effect in long-term cognition.

Even with all the limitations discussed above, the reported estimates are stable and consistent across all specifications tested, and the message is quite clear: preschool attendance matters for cognitive skill formation in the childhood, and since cognition is persistent across stages, investment has an indirect cumulative effect in long-term cognition.

References

ACOSTA, P.; MULLER, N.; SARZOSA, M. A. Beyond qualifications: returns to cognitive and socio-emotional skills in colombia. *World Bank Policy Research Working Paper*, n. 7430, 2015.

ARCIDIACONO, P.; JONES, J. B. Finite mixture distributions, sequential likelihood and the em algorithm. *Econometrica*, Wiley Online Library, v. 71, n. 3, p. 933–946, 2003.

ATTANASIO, O. et al. *Estimating the production function for human capital: Results from a randomized control trial in Colombia*. [S.l.], 2015.

ATTANASIO, O.; MEGHIR, C.; NIX, E. *Human capital development and parental investment in india*. [S.l.], 2015.

ATTANASIO, O. et al. Human capital growth and poverty: Evidence from ethiopia and peru. *Review of Economic Dynamics*, Elsevier, v. 25, p. 234–259, 2017.

BARROS, R. P. de et al. Uma avaliação do impacto da qualidade da creche no desenvolvimento infantil. *Pesquisa e planejamento econômico*, v. 41, p. 213–232, 2011.

BORGHANS, L. et al. The economics and psychology of personality traits. *Journal of human Resources*, University of Wisconsin Press, v. 43, n. 4, p. 972–1059, 2008.

CAMPBELL, F. A. et al. Early childhood education: Young adult outcomes from the abecedarian project. *Applied developmental science*, Taylor & Francis, v. 6, n. 1, p. 42–57, 2002.

CAPRARA, G. V. et al. The "big five questionnaire": A new questionnaire to assess the five factor model. *Personality and individual Differences*, Elsevier, v. 15, n. 3, p. 281–288, 1993.

CARNEIRO, P.; HANSEN, K. T.; HECKMAN, J. J. *Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College*. [S.l.], 2003.

COBB-CLARK, D. A.; SCHURER, S. The stability of big-five personality traits. *Economics Letters*, Elsevier, v. 115, n. 1, p. 11–15, 2012.

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *psychometrika*, Springer, v. 16, n. 3, p. 297–334, 1951.

CUNHA, F.; HECKMAN, J. *The technology of skill formation*. [S.l.], 2007.

CUNHA, F.; HECKMAN, J. J. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of human resources*, University of Wisconsin Press, v. 43, n. 4, p. 738–782, 2008.

CUNHA, F.; HECKMAN, J. J. *Investing in our young people*. [S.l.], 2010.

CUNHA, F.; HECKMAN, J. J.; SCHENNACH, S. M. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, Wiley Online Library, v. 78, n. 3, p. 883–931, 2010.

CURI, A. Z.; MENEZES-FILHO, N. A. et al. Os efeitos da pré-escola sobre os salários, a escolaridade e a proficiência escolar. *Anais da ANPEC*, 2006.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, JSTOR, p. 1–38, 1977.

DEYOUNG, C. G. Personality neuroscience and the biology of traits. *Social and Personality Psychology Compass*, Wiley Online Library, v. 4, n. 12, p. 1165–1180, 2010.

DEYOUNG, C. G. et al. Testing predictions from personality neuroscience: Brain structure and the big five. *Psychological science*, Sage Publications Sage CA: Los Angeles, CA, v. 21, n. 6, p. 820–828, 2010.

FELÍCIO, F.; FERNANDES, R. O efeito da qualidade da escola sobre o desempenho escolar: uma avaliação do ensino fundamental no estado de são paulo. *Anais do XXXIII Encontro Nacional de Economia*, 2005.

FELÍCIO, F.; MENEZES, R.; ZOGHBI, A. C. *The effects of early child education on literacy scores using data from a new Brazilian assessment tool*. [S.l.], 2010.

FELICIO, F. d. et al. *Educação infantil e seus efeitos de curto e médio prazos: Levantamento e análise de dados longitudinais*. [S.l.], 2013.

FELÍCIO, F. d.; TERRA, R.; ZOGHBI, A. C. The effects of early childhood education on literacy scores using data from a new brazilian assessment tool. *Estudos Econômicos (São Paulo)*, Instituto de Pesquisas Econômicas da FEA-USP, v. 42, n. 1, p. 97–128, 2012.

FELÍCIO, F. de; VASCONCELLOS, L. et al. O efeito da educação infantil sobre o desempenho escolar medido em exames padronizados. In: ANPEC-ASSOCIAÇÃO NACIONAL DOS CENTROS DE PÓSGRADUAÇÃO EM ECONOMIA [BRAZILIAN ASSOCIATION OF GRADUATE PROGRAMS IN ECONOMICS]. *Anais do XXXV Encontro Nacional de Economia [Proceedings of the 35th Brazilian Economics Meeting]*. [S.l.], 2007.

FISCHER, R. Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in jccp. *Journal of Cross-Cultural Psychology*, Sage Publications, v. 35, n. 3, p. 263–282, 2004.

FONSECA, G. do C. *Investigação da durabilidade do benefício gerado pela Educação Infantil*. 2015.

HECKMAN, J. The technology and neuroscience of skill formation. *Invest in Kids Working Group, Center for Economic Development, Partnership for America's Economic Success, July*, v. 17, 2006.

HECKMAN, J. J. et al. The rate of return to the highscope perry preschool program. *Journal of public Economics*, Elsevier, v. 94, n. 1-2, p. 114–128, 2010.

HECKMAN, J. J.; STIXRUD, J.; URZUA, S. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, The University of Chicago Press, v. 24, n. 3, p. 411–482, 2006.

PHILLIPS, D. A.; SHONKOFF, J. P. et al. *From neurons to neighborhoods: The science of early childhood development*. [S.l.]: National Academies Press, 2000.

PINTO, C. C. d. X.; SANTOS, D.; GUIMARÃES, C. The impact of daycare attendance on math test scores for a cohort of fourth graders in brazil. *The Journal of Development Studies*, Routledge, v. 53, n. 9, p. 1335–1357, 2017.

PROCÓPIO, I. V. et al. Dois ensaios sobre os determinantes da desigualdade educacional brasileira a partir de dados longitudinais. Universidade Federal de Juiz de Fora, 2012.

RANTANEN, J. et al. Long-term stability in the big five personality traits in adulthood. *Scandinavian Journal of Psychology*, Wiley Online Library, v. 48, n. 6, p. 511–518, 2007.

SANTOS, D. D. dos. *Impactos do ensino infantil sobre o aprendizado: benefícios positivos, mas desiguais*. Tese (Doutorado), 2016.

SOLDZ, S.; VAILLANT, G. E. The big five personality traits and the life course: A 45-year longitudinal study. *Journal of Research in Personality*, Elsevier, v. 33, n. 2, p. 208–232, 1999.

TODD, P. E.; WOLPIN, K. I. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, Wiley Online Library, v. 113, n. 485, 2003.

TRAIN, K. E. *Discrete choice methods with simulation*. [S.l.]: Cambridge university press, 2009.

ULYSSEA, G. L. de. *Essays on human capital formation from gestation to adolescence*. Tese (Doutorado) — PUC-Rio, 2017.

VERAMENDI, G.; URZÚA, S. *The Impact of Out-of-Home Childcare Centers on Early Childhood Development*. [S.l.], 2011.

Appendix

A – The sensibility of the investment variable

In order to build a continuous variable for early childhood investment, the weights to daycare (DC) and kindergarten (KD) years were stablished *ad hoc* in equation 5.7. To verify how our estimates could be sensible to this particular specification of the investment variable, we run several estimates of the synthetic dataset, with different values of the investment variable, and verify how investment and CES equations coefficients would be affected.

Our baseline specification is equivalent to setting $b = 1$ in the equation below:

$$I_{i,1} = \frac{1}{6}HC_i + \left(1 - \frac{b}{6}\right)DC_{0-2i} + \left(1 + \frac{b}{6}\right)(DC_{3i} + KD_i) \tag{A.1}$$

By increasing the value of b we get concave investment variables above the default specification, that assign progressively greater values for kindergarten years relative to daycare years, as shown in figure 10. By decreasing b, we have the investment curves below the default specification, that attribute higher values for daycare years in comparison to kindergarten.

In short, we vary the relative importance of daycare and kindergarten years, such that $I_{i,1}$ remains increasing in the number of preschool years attended, such that the extremes (the investment corresponding to 0 years of preschool and 6 years of daycare+kindergarten) remain unchanged - even though our data is truncated and we do not observe the last two points.

For each value of the adjustment parameter b, we estimate a different synthetic dataset using the corresponding investment variable, keeping all the other variables of the baseline specification. There are no changes in the cognitive and non-cognitive latent factors drawn from those synthetic datasets, neither in the factor loadings of the measures.

In table 11 we report the results of the investment equation of the baseline specification along with the minimum and maximum coefficients obtained using different investment variables, with the adjustment parameter ranging from $b = -4.5$ to $b = 4.5$. The instrument remains negative for all the investment variables tested, being the minimum value achieved in all equations run equal to -0.048 and the maximum -0.042. Income and parent's education remain as important determinants for preschool investment. The variables did not change sign, except by the variable for being seriously ill in early childhood, that is small and negative in the baseline and assumed small positive values in some of the specifications tested. The full table with coefficients obtained from every

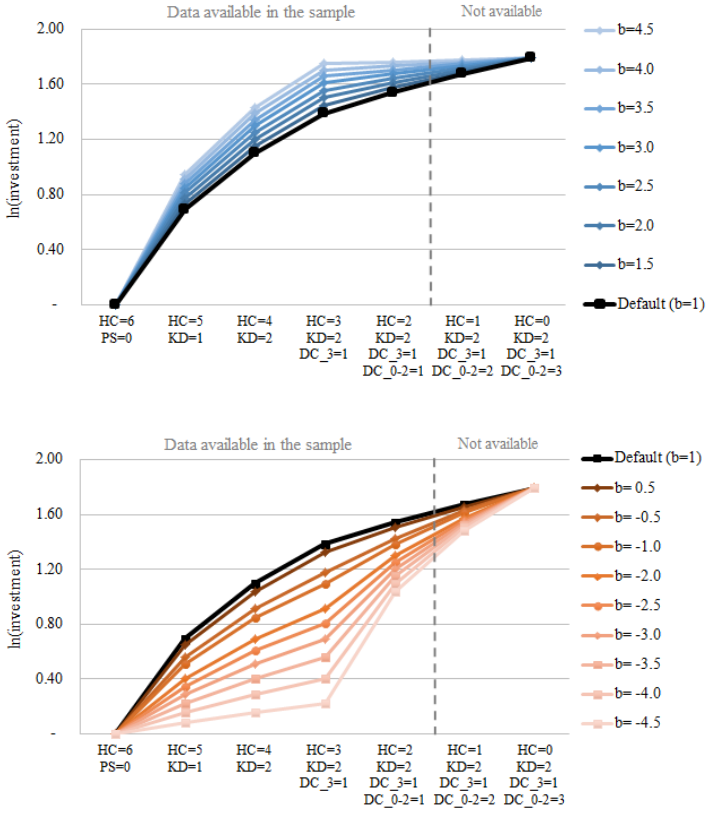


Figure 10: Alternative specifications for the investment variable (in logs)

synthetic database used in this exercise are reported in table 13.

For the CES equation, varying the scale of the investment variable did not change our results either. The minimum and maximum estimates of the share parameters among all specifications are pretty close to our baseline equation, and we still get estimates of the elasticity of substitution close to 1 for every scale for investment we consider. All the estimates with and without control function are reported in tables 14 and 15, respectively.

Table 11: Investment equation comparison: baseline *versus* min. and max. coefficients of alternative specifications for investment

Investment equation coefficients	Baseline	Min.	Max.
Instrument (smallest distance to ECC)	-0.048	-0.048	-0.042
	[-0.1;0.01]		
Socioeconomic Index (SES)	0.116	0.076	0.116
	[0;0.18]		
Male	-0.028	-0.043	-0.028
	[-0.08;0.04]		
White	0.023	0.009	0.023
	[-0.05;0.07]		
Number of siblings	-0.096	-0.096	-0.051
	[-0.14;0]		
Parents' highest education	0.148	0.148	0.203
	[0.09;0.26]		
Parents married in early childhood	-0.047	-0.053	-0.034
	[-0.11;0.03]		
Seriously ill in early childhood	-0.019	-0.044	0.005
	[-0.08;0.04]		
Age	-0.052	-0.058	-0.042
	[-0.1;0.03]		

Note: 95% confidence intervals obtained using 500 bootstrap replications in square brackets

Table 12: CES comparison: baseline *versus* min. and max. coefficients of alternative specifications of the investment variable

CES equation coefficients	Baseline	Min.	Max.	Baseline, no CF	Min.	Max.
Investment (γ)	0.715	0.715	0.739	0.289	0.282	0.300
	[0.64;0.84]			[0.24;0.37]		
Non-cognitive skills ($1-\gamma$)	0.285	0.261	0.285	0.711	0.700	0.718
	[0.16;0.36]			[0.63;0.76]		
Complementarity (ρ)	0.013	0.013	0.041	0.009	0.003	0.051
	[-0.29;0.09]			[-0.14;0.08]		
Elasticity Substitution (σ)	1.013	1.013	1.042	1.009	1.003	1.054
	[0.77;1.1]			[0.88;1.09]		
Control Function (CF)	-0.668	-0.721	-0.668			
	[-0.82;-0.59]					
Intercept	-0.001	-0.005	0.008	-0.002	-0.002	0.009
	[-0.02;0.04]			[-0.02;0.03]		
Socioeconomic Index (SES)	-0.069	-0.069	-0.047	-0.015	-0.015	0.000
	[-0.16;0.05]			[-0.11;0.08]		
Male	-0.088	-0.088	-0.078	-0.026	-0.026	0.000
	[-0.16;-0.02]			[-0.09;0.05]		
White	0.065	0.065	0.071	0.035	-0.001	0.040
	[0.01;0.15]			[-0.03;0.11]		
Age	-0.028	-0.049	-0.018	-0.055	-0.073	0.000
	[-0.13;0.04]			[-0.15;0]		
Number of siblings	-0.045	-0.079	-0.045	-0.060	-0.067	0.000
	[-0.13;0.02]			[-0.14;0.02]		
Parents' highest education	0.085	0.053	0.085	0.134	-0.001	0.134
	[-0.04;0.17]			[0.03;0.21]		

Note: 95% confidence intervals obtained using 500 bootstrap replications in square brackets

Table 13: Investment equation coefficients for all alternative specifications of the investment variable

Investment equation coefficients	b = -4.5	b = -4	b = -3.5	b = -3	b = -2.5	b = -2	b = -1.5	b = -1	b = -0.5
Instrument (smallest distance to ECC	-0.043	-0.043	-0.043	-0.043	-0.043	-0.043	-0.043	-0.042	-0.042
	[-0.1:0.02]	[-0.1:0.02]	[-0.1:0.02]	[-0.1:0.02]	[-0.1:0.02]	[-0.1:0.02]	[-0.11:0.02]	[-0.1:0.02]	[-0.11:0.02]
Socioeconomic Index (SES)	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.090	0.090
	[0:0.17]	[0:0.17]	[0.02:0.18]	[0:0.17]	[0:0.17]	[0:0.17]	[-0.01:0.17]	[-0.01:0.16]	[-0.01:0.16]
Male	-0.033	-0.033	-0.033	-0.033	-0.033	-0.033	-0.033	-0.043	-0.043
	[-0.08:0.04]	[-0.08:0.04]	[-0.06:0.07]	[-0.08:0.04]	[-0.08:0.04]	[-0.08:0.04]	[-0.09:0.03]	[-0.09:0.03]	[-0.09:0.03]
White	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.022	0.022
	[-0.05:0.07]	[-0.05:0.07]	[-0.03:0.11]	[-0.05:0.07]	[-0.05:0.07]	[-0.05:0.07]	[-0.05:0.07]	[-0.05:0.07]	[-0.05:0.07]
Number of siblings	-0.075	-0.075	-0.075	-0.075	-0.075	-0.075	-0.075	-0.051	-0.051
	[-0.15:-0.01]	[-0.15:-0.01]	[-0.14:-0.01]	[-0.15:-0.01]	[-0.15:-0.01]	[-0.15:-0.01]	[-0.13:0.02]	[-0.12:0.02]	[-0.12:0.02]
Parents' highest education	0.170	0.170	0.170	0.170	0.170	0.170	0.170	0.173	0.173
	[0.09:0.26]	[0.09:0.26]	[0.06:0.24]	[0.09:0.26]	[0.09:0.26]	[0.09:0.26]	[0.09:0.26]	[0.08:0.26]	[0.09:0.26]
Parents married in early childhood	-0.040	-0.040	-0.040	-0.040	-0.040	-0.040	-0.040	-0.034	-0.034
	[-0.1:0.02]	[-0.1:0.02]	[-0.07:0.05]	[-0.1:0.02]	[-0.1:0.02]	[-0.1:0.02]	[-0.1:0.03]	[-0.1:0.03]	[-0.1:0.03]
Seriously ill in early childhood	-0.035	-0.035	-0.035	-0.035	-0.035	-0.035	-0.035	-0.044	-0.044
	[-0.08:0.05]	[-0.08:0.05]	[-0.11:0.02]	[-0.08:0.05]	[-0.08:0.05]	[-0.08:0.05]	[-0.09:0.04]	[-0.09:0.04]	[-0.09:0.04]
Age	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052	-0.052	-0.058	-0.058
	[-0.1:0.03]	[-0.1:0.03]	[-0.05:0.08]	[-0.1:0.03]	[-0.1:0.03]	[-0.1:0.03]	[-0.11:0.02]	[-0.11:0.02]	[-0.11:0.02]

Investment equation coefficients	b = 0.5	b = 1	b = 1.5	b = 2	b = 2.5	b = 3	b = 3.5	b = 4	b = 4.5
Instrument (smallest distance to ECC	-0.043	-0.048	-0.042	-0.044	-0.044	-0.044	-0.044	-0.044	-0.044
	[-0.1:0.02]	[-0.1;0.01]	[-0.1:0.01]	[-0.11:0.02]	[-0.11:0.02]	[-0.11:0.02]	[-0.11:0.02]	[-0.11:0.02]	[-0.11:0.02]
Socioeconomic Index (SES)	0.097	0.116	0.076	0.100	0.100	0.100	0.101	0.101	0.101
	[0:0.17]	[0;0.18]	[0.01:0.17]	[0.01:0.17]	[0.01:0.17]	[0.01:0.18]	[0.01:0.18]	[0.01:0.18]	[0.01:0.18]
Male	-0.033	-0.028	-0.036	-0.033	-0.032	-0.032	-0.032	-0.032	-0.032
	[-0.08:0.04]	[-0.08;0.04]	[-0.09:0.04]	[-0.08:0.04]	[-0.08:0.03]	[-0.08:0.04]	[-0.08:0.04]	[-0.08:0.04]	[-0.08:0.04]
White	0.018	0.023	0.009	0.018	0.018	0.018	0.018	0.018	0.018
	[-0.05:0.07]	[-0.05;0.07]	[-0.06:0.06]	[-0.05:0.07]	[-0.06:0.07]	[-0.05:0.07]	[-0.05:0.07]	[-0.05:0.07]	[-0.05:0.07]
Number of siblings	-0.074	-0.096	-0.072	-0.078	-0.078	-0.079	-0.080	-0.080	-0.081
	[-0.15:-0.01]	[-0.14;0]	[-0.15:0]	[-0.15:-0.01]	[-0.15:-0.02]	[-0.15:-0.01]	[-0.15:-0.01]	[-0.15:-0.01]	[-0.15:-0.01]
Parents' highest education	0.171	0.148	0.203	0.169	0.169	0.168	0.168	0.167	0.167
	[0.09:0.26]	[0.09;0.26]	[0.1:0.27]	[0.09:0.26]	[0.09:0.25]	[0.09:0.26]	[0.09:0.26]	[0.09:0.26]	[0.09:0.26]
Parents married in early childhood	-0.041	-0.047	-0.053	-0.038	-0.037	-0.037	-0.036	-0.036	-0.035
	[-0.11:0.02]	[-0.11;0.03]	[-0.11:0.02]	[-0.1:0.02]	[-0.1:0.02]	[-0.1:0.02]	[-0.1:0.02]	[-0.1:0.02]	[-0.1:0.03]
Seriously ill in early childhood	-0.035	-0.019	0.005	-0.035	-0.035	-0.035	-0.035	-0.035	-0.035
	[-0.08:0.05]	[-0.08;0.04]	[-0.08:0.04]	[-0.08:0.05]	[-0.08:0.05]	[-0.08:0.05]	[-0.08:0.05]	[-0.08:0.05]	[-0.08:0.05]
Age	-0.051	-0.052	-0.042	-0.053	-0.053	-0.053	-0.054	-0.054	-0.054
	[-0.1:0.03]	[-0.1;0.03]	[-0.11:0.02]	[-0.1:0.03]	[-0.1:0.03]	[-0.1:0.03]	[-0.1:0.03]	[-0.1:0.03]	[-0.1:0.03]

Note: each column was generated from a different synthetic database; b=1 corresponds to the baseline equation

Table 14: CES equation coefficients (with CF) for all alternative specifications of the investment variable

CES equation coefficients	b = -4.5	b = -4	b = -3.5	b = -3	b = -2.5	b = -2	b = -1.5	b = -1	b = -0.5
Investment (γ)	0.737	0.737	0.737	0.737	0.737	0.737	0.737	0.737	0.737
	[0.64:0.84]	[0.64:0.84]	[0.62:0.82]	[0.64:0.84]	[0.64:0.84]	[0.64:0.84]	[0.64:0.84]	[0.64:0.84]	[0.64:0.84]
Non-cognitive skills (1-γ)	0.263	0.263	0.263	0.263	0.263	0.263	0.263	0.263	0.263
	[0.16:0.36]	[0.16:0.36]	[0.18:0.38]	[-0.02:0.04]	[0.16:0.36]	[0.16:0.36]	[0.16:0.36]	[0.16:0.36]	[0.16:0.36]
Complementarity (ρ)	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.020	0.020
	[-0.28:0.11]	[-0.28:0.11]	[-0.14:0.11]	[-0.15:0.04]	[-0.28:0.11]	[-0.28:0.11]	[-0.26:0.1]	[-0.25:0.11]	[-0.24:0.1]
Elasticity Substitution (σ)	1.016	1.016	1.016	1.016	1.016	1.016	1.016	1.020	1.020
	[0.78:1.12]	[0.78:1.12]	[0.87:1.12]	[-0.16:-0.02]	[0.78:1.12]	[0.78:1.12]	[0.79:1.11]	[0.8:1.12]	[0.8:1.11]
Control Function (CF)	-0.696	-0.696	-0.696	-0.696	-0.696	-0.696	-0.696	-0.721	-0.721
	[-0.81:-0.59]	[-0.81:-0.59]	[-0.78:-0.56]	[-0.81:-0.59]	[-0.81:-0.59]	[-0.81:-0.59]	[-0.83:-0.6]	[-0.83:-0.61]	[-0.83:-0.61]
Intercept	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.005	-0.005
	[-0.02:0.04]	[-0.02:0.04]	[-0.03:0.03]	[-0.02:0.04]	[-0.02:0.04]	[-0.02:0.04]	[-0.02:0.04]	[-0.02:0.04]	[-0.02:0.04]
Socioeconomic Index (SES)	-0.057	-0.057	-0.057	-0.057	-0.057	-0.057	-0.057	-0.052	-0.052
	[-0.15:0.04]	[-0.15:0.04]	[-0.21:-0.02]	[-0.15:0.04]	[-0.15:0.04]	[-0.15:0.04]	[-0.15:0.04]	[-0.15:0.05]	[-0.15:0.05]
Male	-0.087	-0.087	-0.087	-0.087	-0.087	-0.087	-0.087	-0.079	-0.079
	[-0.16:-0.02]	[-0.16:-0.02]	[-0.16:-0.01]	[-0.16:-0.02]	[-0.16:-0.02]	[-0.16:-0.02]	[-0.16:-0.01]	[-0.16:-0.01]	[-0.16:-0.01]
White	0.071	0.071	0.071	0.071	0.071	0.071	0.071	0.068	0.068
	[0.01:0.15]	[0.01:0.15]	[0:0.14]	[0.01:0.15]	[0.01:0.15]	[0.01:0.15]	[0.01:0.15]	[0:0.15]	[0.01:0.15]
Age	-0.023	-0.023	-0.023	-0.023	-0.023	-0.023	-0.023	-0.018	-0.018
	[-0.13:0.03]	[-0.13:0.03]	[-0.11:0.06]	[-0.13:0.03]	[-0.13:0.03]	[-0.13:0.03]	[-0.12:0.03]	[-0.12:0.03]	[-0.12:0.03]
Number of siblings	-0.062	-0.062	-0.062	-0.062	-0.062	-0.062	-0.062	-0.079	-0.079
	[-0.13:0.03]	[-0.13:0.03]	[-0.13:0.03]	[-0.13:0.03]	[-0.13:0.03]	[-0.13:0.03]	[-0.15:0.01]	[-0.16:0.01]	[-0.16:0.01]
Parents' highest education	0.069	0.069	0.069	0.069	0.069	0.069	0.069	0.067	0.067
	[-0.05:0.17]	[-0.05:0.17]	[0.03:0.21]	[-0.05:0.17]	[-0.05:0.17]	[-0.05:0.17]	[-0.05:0.17]	[-0.05:0.16]	[-0.05:0.17]

CES equation coefficients	b = 0.5	b = 1	b = 1.5	b = 2	b = 2.5	b = 3	b = 3.5	b = 4	b = 4.5
Investment (γ)	0.737	0.715	0.739	0.738	0.738	0.738	0.738	0.738	0.739
	[0.64:0.84]	[0.64;0.84]	[0.64:0.84]	[0.64:0.84]	[0.63:0.84]	[0.64:0.84]	[0.64:0.84]	[0.64:0.84]	[0.64:0.84]
Non-cognitive skills (1-γ)	0.263	0.285	0.261	0.262	0.262	0.262	0.262	0.262	0.261
	[0.16:0.36]	[0.16;0.36]	[0.16:0.36]	[0.16:0.36]	[0.16:0.37]	[0.16:0.36]	[0.16:0.36]	[0.16:0.36]	[0.16:0.36]
Complementarity (ρ)	0.016	0.013	0.041	0.016	0.016	0.016	0.016	0.016	0.016
	[-0.26:0.11]	[-0.29;0.09]	[-0.23:0.09]	[-0.31:0.11]	[-0.29:0.1]	[-0.32:0.11]	[-0.37:0.11]	[-0.39:0.11]	[-0.39:0.11]
Elasticity Substitution (σ)	1.016	1.013	1.042	1.016	1.016	1.016	1.016	1.016	1.016
	[0.79:1.12]	[0.77;1.1]	[0.82:1.1]	[0.77:1.12]	[0.78:1.11]	[0.76:1.12]	[0.73:1.12]	[0.72:1.13]	[0.72:1.13]
Control Function (CF)	-0.696	-0.668	-0.701	-0.697	-0.697	-0.698	-0.698	-0.698	-0.699
	[-0.81:-0.59]	[-0.82;-0.59]	[-0.8:-0.58]	[-0.81:-0.59]	[-0.81:-0.59]	[-0.81:-0.59]	[-0.81:-0.58]	[-0.81:-0.58]	[-0.81:-0.58]
Intercept	-0.004	-0.001	0.008	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004
	[-0.02:0.04]	[-0.02;0.04]	[-0.03:0.04]	[-0.02:0.04]	[-0.02:0.04]	[-0.02:0.05]	[-0.02:0.05]	[-0.02:0.06]	[-0.02:0.06]
Socioeconomic Index (SES)	-0.056	-0.069	-0.047	-0.058	-0.059	-0.059	-0.060	-0.060	-0.060
	[-0.15:0.04]	[-0.16;0.05]	[-0.16:0.03]	[-0.16:0.04]	[-0.15:0.04]	[-0.16:0.04]	[-0.16:0.04]	[-0.16:0.04]	[-0.16:0.04]
Male	-0.086	-0.088	-0.078	-0.087	-0.087	-0.087	-0.088	-0.088	-0.088
	[-0.16:-0.02]	[-0.16;-0.02]	[-0.15:-0.02]	[-0.16:-0.02]	[-0.16:-0.01]	[-0.16:-0.02]	[-0.16:-0.02]	[-0.16:-0.02]	[-0.16:-0.02]
White	0.071	0.065	0.065	0.071	0.071	0.071	0.071	0.071	0.071
	[0.01:0.15]	[0.01;0.15]	[0.01:0.16]	[0.01:0.15]	[0.01:0.15]	[0.01:0.15]	[0.01:0.15]	[0.01:0.15]	[0.01:0.15]
Age	-0.023	-0.028	-0.049	-0.022	-0.022	-0.022	-0.021	-0.021	-0.021
	[-0.13:0.03]	[-0.13;0.04]	[-0.13:0.03]	[-0.12:0.03]	[-0.13:0.03]	[-0.12:0.03]	[-0.12:0.03]	[-0.12:0.03]	[-0.12:0.03]
Number of siblings	-0.063	-0.045	-0.051	-0.060	-0.060	-0.059	-0.059	-0.058	-0.058
	[-0.13:0.02]	[-0.13;0.02]	[-0.14:0.02]	[-0.14:0.03]	[-0.14:0.02]	[-0.13:0.03]	[-0.13:0.03]	[-0.13:0.03]	[-0.13:0.03]
Parents' highest education	0.069	0.085	0.053	0.070	0.071	0.071	0.071	0.072	0.072
	[-0.05:0.16]	[-0.04;0.17]	[-0.04:0.17]	[-0.05:0.17]	[-0.04:0.16]	[-0.05:0.17]	[-0.04:0.17]	[-0.04:0.17]	[-0.04:0.17]

Note: each column was generated from a different synthetic database; b=1 corresponds to the baseline equation

Table 15: CES equation coefficients (without CF) for all alternative specifications of the investment variable

CES equation coefficients, no CF	b = -4.5	b = -4	b = -3.5	b = -3	b = -2.5	b = -2	b = -1.5	b = -1	b = -0.5
Investment (γ)	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.282	0.282
	[0.24:0.38]	[0.24:0.38]	[0.21:0.38]	[0.24:0.38]	[0.24:0.38]	[0.24:0.38]	[0.22:0.37]	[0.22:0.36]	[0.22:0.36]
Non-cognitive skills (1-γ)	0.700	0.700	0.700	0.700	0.700	0.700	0.700	0.718	0.718
	[0.62:0.76]	[0.62:0.76]	[0.62:0.79]	[-0.02:0.03]	[0.62:0.76]	[0.62:0.76]	[0.63:0.78]	[0.64:0.78]	[0.64:0.78]
Complementarity (ρ)	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.005	0.005
	[-0.14:0.1]	[-0.14:0.1]	[-0.11:0.09]	[-0.11:0.08]	[-0.14:0.1]	[-0.14:0.1]	[-0.14:0.1]	[-0.14:0.1]	[-0.12:0.1]
Elasticity Substitution (σ)	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.005	1.005
	[0.88:1.11]	[0.88:1.11]	[0.9:1.1]	[-0.09:0.05]	[0.88:1.11]	[0.88:1.11]	[0.88:1.11]	[0.88:1.11]	[0.89:1.11]
Intercept	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.001	-0.001
	[-0.02:0.03]	[-0.02:0.03]	[-0.02:0.03]	[-0.02:0.03]	[-0.02:0.03]	[-0.02:0.03]	[-0.02:0.03]	[-0.02:0.03]	[-0.02:0.03]
Socioeconomic Index (SES)	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	0.000	0.000
	[-0.11:0.08]	[-0.11:0.08]	[-0.16:0.02]	[-0.11:0.08]	[-0.11:0.08]	[-0.11:0.08]	[-0.11:0.08]	[-0.11:0.08]	[-0.11:0.08]
Male	-0.023	-0.023	-0.023	-0.023	-0.023	-0.023	-0.023	-0.017	-0.017
	[-0.09:0.05]	[-0.09:0.05]	[-0.08:0.06]	[-0.09:0.05]	[-0.09:0.05]	[-0.09:0.05]	[-0.08:0.06]	[-0.08:0.06]	[-0.09:0.06]
White	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.037	0.037
	[-0.03:0.11]	[-0.03:0.11]	[-0.02:0.12]	[-0.03:0.11]	[-0.03:0.11]	[-0.03:0.11]	[-0.03:0.11]	[-0.03:0.11]	[-0.03:0.11]
Age	-0.050	-0.050	-0.050	-0.050	-0.050	-0.050	-0.050	-0.049	-0.049
	[-0.14:0]	[-0.14:0]	[-0.08:0.07]	[-0.14:0]	[-0.14:0]	[-0.14:0]	[-0.15:0]	[-0.15:0]	[-0.15:0]
Number of siblings	-0.067	-0.067	-0.067	-0.067	-0.067	-0.067	-0.067	-0.074	-0.074
	[-0.14:0.02]	[-0.14:0.02]	[-0.14:0.02]	[-0.14:0.02]	[-0.14:0.02]	[-0.14:0.02]	[-0.14:0.02]	[-0.15:0.01]	[-0.15:0.02]
Parents' highest education	0.124	0.124	0.124	0.124	0.124	0.124	0.124	0.126	0.126
	[0.03:0.22]	[0.03:0.22]	[0.08:0.26]	[0.03:0.22]	[0.03:0.22]	[0.03:0.22]	[0.03:0.22]	[0.03:0.22]	[0.03:0.22]

CES equation coefficients, no CF	b = 0.5	b = 1	b = 1.5	b = 2	b = 2.5	b = 3	b = 3.5	b = 4	b = 4.5
Investment (γ)	0.300	0.289	0.295	0.300	0.300	0.299	0.299	0.299	0.299
	[0.24:0.38]	[0.24;0.37]	[0.23:0.38]	[0.24:0.38]	[0.23:0.38]	[0.24:0.38]	[0.24:0.38]	[0.24:0.38]	[0.24:0.38]
Non-cognitive skills (1-γ)	0.700	0.711	0.705	0.700	0.700	0.701	0.701	0.701	0.701
	[0.62:0.76]	[0.63;0.76]	[0.62:0.77]	[0.62:0.76]	[0.62:0.77]	[0.62:0.76]	[0.62:0.76]	[0.62:0.76]	[0.62:0.76]
Complementarity (ρ)	0.003	0.009	0.051	0.003	0.003	0.003	0.003	0.003	0.003
	[-0.13:0.1]	[-0.14;0.08]	[-0.12:0.09]	[-0.13:0.1]	[-0.16:0.09]	[-0.12:0.11]	[-0.12:0.11]	[-0.12:0.11]	[-0.12:0.11]
Elasticity Substitution (σ)	1.003	1.009	1.054	1.003	1.003	1.003	1.003	1.003	1.003
	[0.88:1.11]	[0.88;1.09]	[0.9:1.09]	[0.89:1.11]	[0.86:1.1]	[0.89:1.12]	[0.89:1.12]	[0.89:1.12]	[0.89:1.13]
Intercept	0.000	-0.002	0.009	0.000	0.000	0.000	0.000	0.000	0.000
	[-0.02:0.03]	[-0.02;0.03]	[-0.03:0.03]	[-0.02:0.03]	[-0.02:0.03]	[-0.02:0.03]	[-0.02:0.03]	[-0.02:0.03]	[-0.03:0.03]
Socioeconomic Index (SES)	-0.004	-0.015	-0.007	-0.004	-0.004	-0.005	-0.005	-0.005	-0.005
	[-0.11:0.08]	[-0.11;0.08]	[-0.12:0.07]	[-0.11:0.08]	[-0.11:0.08]	[-0.11:0.08]	[-0.11:0.08]	[-0.11:0.07]	[-0.11:0.07]
Male	-0.023	-0.026	-0.017	-0.023	-0.023	-0.023	-0.023	-0.023	-0.023
	[-0.09:0.05]	[-0.09;0.05]	[-0.09:0.04]	[-0.09:0.05]	[-0.08:0.04]	[-0.09:0.05]	[-0.09:0.05]	[-0.09:0.05]	[-0.09:0.05]
White	0.040	0.035	0.030	0.040	0.040	0.040	0.040	0.040	0.040
	[-0.03:0.11]	[-0.03;0.11]	[-0.02:0.12]	[-0.03:0.11]	[-0.03:0.11]	[-0.03:0.11]	[-0.03:0.11]	[-0.03:0.11]	[-0.03:0.11]
Age	-0.051	-0.055	-0.073	-0.050	-0.050	-0.050	-0.050	-0.050	-0.050
	[-0.14:0]	[-0.15;0]	[-0.16:0.02]	[-0.14:0]	[-0.14:0]	[-0.14:0]	[-0.14:0]	[-0.14:0]	[-0.14:0]
Number of siblings	-0.067	-0.060	-0.053	-0.066	-0.066	-0.065	-0.065	-0.065	-0.065
	[-0.14:0.02]	[-0.14;0.02]	[-0.14:0]	[-0.14:0.02]	[-0.14:0.02]	[-0.14:0.02]	[-0.14:0.02]	[-0.14:0.02]	[-0.14:0.02]
Parents' highest education	0.124	0.134	0.122	0.125	0.125	0.125	0.125	0.125	0.126
	[0.03:0.22]	[0.03;0.21]	[0.04:0.22]	[0.03:0.22]	[0.03:0.2]	[0.03:0.22]	[0.03:0.22]	[0.03:0.22]	[0.03:0.22]

Note: each column was generated from a different synthetic database; b=1 corresponds to the baseline equation

B – Results for alternative specifications

A summary of all results reported in the sections B.1 - B.4 is presented in the following tables 16 to 20.

Table 16: Factor Loadings summary - Baseline and Appendix B.1- B.4

		Baseline		Corrected for Aquiescence		Without Literacy (Provinha)		Reduced Sample (706 obs)	
Latent Factor	Measures	Loading	Signal	Loading	Signal	Loading	Signal	Loading	Signal
Cognitive skills	Portuguese Score	1.00	89.5%	1.00	88.7%	1.00	97.2%	1.00	85.7%
	Math Score	0.86	68.7%	0.88	70.6%	0.76	55.3%	0.85	63.4%
	Literacy Score	0.57	31.1%	0.56	30.1%			0.62	38.7%
Non-cognitive skills	Conscientiousness	1.00	60.6%	1.00	52.3%	1.00	60.0%	1.00	57.1%
	Neurocitism	0.90	49.4%	0.93	44.3%	0.91	50.0%	0.90	47.8%
	Openness	1.00	59.0%	1.13	65.2%	1.00	58.5%	1.03	60.0%
	Agreeableness	1.04	62.6%	1.11	59.9%	1.06	63.6%	1.06	63.4%
	Extraversion	0.91	48.9%	0.92	42.4%	0.91	48.4%	0.90	45.5%

Social Skills

Latent Factor	Measures	Loading	Signal
Cognitive skills	Portuguese Score	1.00	79.6%
	Math Score	0.93	72.4%
	Literacy Score	0.63	35.2%
Non-cognitive skills	Cooperation	1.00	63.5%
	Assertiveness	1.01	66.6%
	Empathy	0.92	51.5%
	Self-control	0.94	57.9%

Table 17: Investment equation coefficients summary - Baseline and Appendix B.1- B.4

Investment equation coefficients	Baseline	Social Skills Measures	Corrected for Aquiescence	Without Literacy	Reduced Sample (706 obs)
Instrument (smallest distance to ECC)	-0.048	-0.043	-0.043	-0.047	-0.048
	[-0.1 ; 0.01]	[-0.11 ; 0.02]	[-0.1 ; 0.02]	[-0.11 ; 0.01]	[-0.12 ; 0.01]
Socioeconomic Index (SES)	0.116	0.099	0.098	0.078	0.078
	[0 ; 0.18]	[0.01 ; 0.17]	[0.01 ; 0.17]	[-0.01 ; 0.16]	[-0.04 ; 0.16]
Male	-0.028	-0.032	-0.033	-0.025	-0.073
	[-0.08 ; 0.04]	[-0.08 ; 0.04]	[-0.08 ; 0.04]	[-0.08 ; 0.04]	[-0.14 ; 0.01]
White	0.023	0.018	0.018	-0.014	0.020
	[-0.05 ; 0.07]	[-0.05 ; 0.07]	[-0.05 ; 0.07]	[-0.06 ; 0.07]	[-0.06 ; 0.09]
Age	-0.052	-0.052	-0.052	-0.042	-0.067
	[-0.1 ; 0.03]	[-0.1 ; 0.03]	[-0.1 ; 0.03]	[-0.11 ; 0.01]	[-0.14 ; 0.03]
Number of siblings	-0.096	-0.075	-0.075	-0.085	-0.073
	[-0.14 ; 0]	[-0.15 ; -0.01]	[-0.15 ; -0.01]	[-0.15 ; -0.01]	[-0.16 ; 0.01]
Parents' highest education	0.148	0.170	0.170	0.176	0.190
	[0.09 ; 0.26]	[0.09 ; 0.26]	[0.09 ; 0.26]	[0.09 ; 0.27]	[0.09 ; 0.3]
Parents married in early childhood	-0.047	-0.040	-0.040	-0.042	-0.022
	[-0.11 ; 0.03]	[-0.1 ; 0.02]	[-0.1 ; 0.02]	[-0.11 ; 0.03]	[-0.1 ; 0.05]
Seriously ill in early childhood	-0.019	-0.035	-0.035	-0.025	-0.043
	[-0.08 ; 0.04]	[-0.08 ; 0.04]	[-0.08 ; 0.05]	[-0.09 ; 0.05]	[-0.11 ; 0.05]

Note: each column estimates corresponds to a different synthetic dataset.
95% confidence intervals obtained using 500 bootstrap replications (departing from each synthetic dataset) in square brackets.

Table 18: Investment equation coefficients summary, without age - Baseline and Appendix B.1- B.4

Investment equation coefficients (without age)	Baseline	Social Skills Measures	Corrected for Aquiescence	Without Literacy	Reduced Sample (706 obs)
Instrument (smallest distance to ECC)	-0.048	-0.043	-0.042	-0.047	-0.050
	[-0.1 ; 0.01]	[-0.1 ; 0.02]	[-0.1 ; 0.02]	[-0.11 ; 0.01]	[-0.12 ; 0.01]
Socioeconomic Index (SES)	0.112	0.096	0.096	0.075	0.073
	[0 ; 0.17]	[0 ; 0.17]	[0 ; 0.17]	[-0.01 ; 0.16]	[-0.04 ; 0.16]
Male	-0.032	-0.037	-0.037	-0.029	-0.076
	[-0.08 ; 0.03]	[-0.08 ; 0.03]	[-0.08 ; 0.03]	[-0.09 ; 0.04]	[-0.14 ; 0]
White	0.024	0.019	0.019	-0.013	0.024
	[-0.05 ; 0.07]	[-0.05 ; 0.07]	[-0.05 ; 0.07]	[-0.06 ; 0.07]	[-0.06 ; 0.09]
Number of siblings	-0.101	-0.080	-0.080	-0.090	-0.076
	[-0.15 ; -0.01]	[-0.15 ; -0.01]	[-0.15 ; -0.01]	[-0.16 ; -0.01]	[-0.16 ; 0]
Parents' highest education	0.157	0.179	0.179	0.184	0.200
	[0.1 ; 0.26]	[0.1 ; 0.27]	[0.1 ; 0.27]	[0.1 ; 0.27]	[0.1 ; 0.31]
Parents married in early childhood	-0.045	-0.038	-0.038	-0.041	-0.020
	[-0.11 ; 0.03]	[-0.1 ; 0.02]	[-0.1 ; 0.02]	[-0.1 ; 0.03]	[-0.1 ; 0.05]
Seriously ill in early childhood	-0.017	-0.033	-0.033	-0.023	-0.042
	[-0.08 ; 0.05]	[-0.08 ; 0.05]	[-0.08 ; 0.05]	[-0.08 ; 0.05]	[-0.11 ; 0.05]

Note: each column estimates corresponds to a different synthetic dataset.
95% confidence intervals obtained using 500 bootstrap replications (departing from each synthetic dataset) in square brackets.

Table 19: CES equation coefficients summary - Baseline and Appendix B.1- B.4

CES coefficients	Baseline		Social Skills Measures		Corrected for Aquiescence		Without Literacy		Reduced Sample (706 obs)	
	CF	no CF	CF	no CF	CF	no CF	CF	no CF	CF	no CF
Investment (γ)	0.715	0.289	0.809	0.362	0.750	0.294	0.719	0.286	0.761	0.297
	[0.64 ; 0.84]	[0.24 ; 0.37]	[0.72 ; 0.9]	[0.3 ; 0.42]	[0.64 ; 0.87]	[0.22 ; 0.38]	[0.62 ; 0.81]	[0.22 ; 0.36]	[0.62 ; 0.88]	[0.2 ; 0.4]
Non-cognitive skills (1-γ)	0.285	0.711	0.191	0.638	0.250	0.706	0.281	0.714	0.239	0.703
	[0.16 ; 0.36]	[0.63 ; 0.76]	[0.1 ; 0.28]	[0.58 ; 0.7]	[0.13 ; 0.36]	[0.62 ; 0.78]	[0.19 ; 0.38]	[0.64 ; 0.78]	[0.12 ; 0.38]	[0.6 ; 0.8]
Complementarity (ρ)	0.013	0.009	0.017	-0.002	0.029	0.019	-0.045	-0.066	0.018	0.002
	[-0.29 ; 0.09]	[-0.14 ; 0.08]	[-0.19 ; 0.11]	[-0.09 ; 0.08]	[-0.25 ; 0.13]	[-0.12 ; 0.1]	[-0.16 ; 0.09]	[-0.11 ; 0.1]	[-0.42 ; 0.15]	[-0.21 ; 0.12]
Elasticity Substitution (σ)	1.013	1.009	1.017	0.998	1.030	1.019	0.957	0.938	1.019	1.002
	[0.77 ; 1.1]	[0.88 ; 1.09]	[0.84 ; 1.12]	[0.92 ; 1.09]	[0.8 ; 1.14]	[0.89 ; 1.11]	[0.86 ; 1.1]	[0.9 ; 1.11]	[0.7 ; 1.17]	[0.83 ; 1.14]
Control Function (CF)	-0.668		-0.758		-0.700		-0.685		-0.738	
	[-0.82 ; -0.59]		[-0.87 ; -0.64]		[-0.83 ; -0.57]		[-0.78 ; -0.55]		[-0.87 ; -0.58]	
Intercept	-0.001	-0.002	-0.004	0.000	-0.006	-0.002	0.010	0.006	-0.005	-0.001
	[-0.02 ; 0.04]	[-0.02 ; 0.03]	[-0.02 ; 0.03]	[-0.02 ; 0.02]	[-0.02 ; 0.04]	[-0.02 ; 0.03]	[-0.03 ; 0.03]	[-0.03 ; 0.02]	[-0.03 ; 0.05]	[-0.03 ; 0.04]
Socioeconomic Index (SES)	-0.069	-0.015	-0.078	-0.052	-0.059	-0.005	-0.067	-0.023	-0.042	0.029
	[-0.16 ; 0.05]	[-0.11 ; 0.08]	[-0.18 ; 0.02]	[-0.16 ; 0.03]	[-0.16 ; 0.04]	[-0.11 ; 0.09]	[-0.14 ; 0.05]	[-0.11 ; 0.09]	[-0.16 ; 0.08]	[-0.09 ; 0.14]
Male	-0.088	-0.026	-0.117	-0.098	-0.095	-0.040	-0.098	-0.027	-0.004	0.048
	[-0.16 ; -0.02]	[-0.09 ; 0.05]	[-0.19 ; -0.04]	[-0.16 ; -0.02]	[-0.17 ; -0.02]	[-0.11 ; 0.03]	[-0.16 ; -0.02]	[-0.09 ; 0.06]	[-0.09 ; 0.09]	[-0.04 ; 0.15]
White	0.065	0.035	0.079	0.051	0.080	0.062	0.096	0.051	0.030	-0.008
	[0.01 ; 0.15]	[-0.03 ; 0.11]	[0.01 ; 0.16]	[-0.02 ; 0.12]	[0.02 ; 0.15]	[-0.01 ; 0.13]	[0 ; 0.15]	[-0.04 ; 0.11]	[-0.07 ; 0.13]	[-0.1 ; 0.08]
Age	-0.028	-0.055	-0.010	-0.016	-0.017	-0.034	-0.041	-0.067	-0.024	-0.080
	[-0.13 ; 0.04]	[-0.15 ; 0]	[-0.11 ; 0.05]	[-0.12 ; 0.04]	[-0.12 ; 0.04]	[-0.13 ; 0.03]	[-0.13 ; 0.03]	[-0.14 ; 0]	[-0.14 ; 0.06]	[-0.18 ; -0.01]
Number of siblings	-0.045	-0.060	-0.071	-0.098	-0.065	-0.076	-0.043	-0.053	-0.044	-0.058
	[-0.13 ; 0.02]	[-0.14 ; 0.02]	[-0.14 ; 0.02]	[-0.17 ; -0.02]	[-0.14 ; 0.02]	[-0.15 ; 0.02]	[-0.13 ; 0.03]	[-0.14 ; 0.02]	[-0.15 ; 0.06]	[-0.16 ; 0.03]
Parents' highest education	0.085	0.134	0.062	0.123	0.076	0.148	0.071	0.125	0.027	0.092
	[-0.04 ; 0.17]	[0.03 ; 0.21]	[-0.06 ; 0.16]	[0.02 ; 0.22]	[-0.04 ; 0.18]	[0.05 ; 0.25]	[-0.04 ; 0.16]	[0.02 ; 0.22]	[-0.09 ; 0.14]	[-0.01 ; 0.2]

Note: each column estimates corresponds to a different synthetic dataset. Estimates with and without CF run in the same data. CF is the residual of investment equation that includes age. 95% confidence intervals obtained using 500 bootstrap replications in square brackets.

Table 20: CES equation coefficients summary, without age - Baseline and Appendix B.1- B.4

CES coefficients (without age)	Baseline		Social Skills Measures		Corrected for Aquiescence		Without Literacy		Reduced Sample (706 obs)	
	CF	no CF	CF	no CF	CF	no CF	CF	no CF	CF	no CF
Investment (γ)	0.718	0.292	0.809	0.362	0.751	0.295	0.724	0.288	0.766	0.303
	[0.65 ; 0.84]	[0.24 ; 0.38]	[0.72 ; 0.9]	[0.3 ; 0.42]	[0.64 ; 0.87]	[0.22 ; 0.38]	[0.62 ; 0.81]	[0.23 ; 0.36]	[0.63 ; 0.89]	[0.21 ; 0.41]
Non-cognitive skills (1-γ)	0.282	0.708	0.191	0.638	0.249	0.705	0.276	0.712	0.234	0.697
	[0.16 ; 0.35]	[0.62 ; 0.76]	[0.1 ; 0.28]	[0.58 ; 0.7]	[0.13 ; 0.36]	[0.62 ; 0.78]	[0.19 ; 0.38]	[0.64 ; 0.77]	[0.11 ; 0.37]	[0.59 ; 0.79]
Complementarity (ρ)	0.015	0.012	0.016	-0.001	0.030	0.021	-0.047	-0.069	0.020	0.006
	[-0.29 ; 0.09]	[-0.14 ; 0.09]	[-0.19 ; 0.11]	[-0.09 ; 0.08]	[-0.24 ; 0.13]	[-0.12 ; 0.1]	[-0.16 ; 0.09]	[-0.1 ; 0.1]	[-0.44 ; 0.15]	[-0.22 ; 0.12]
Elasticity Substitution (σ)	1.015	1.012	1.016	0.999	1.031	1.022	0.955	0.936	1.020	1.006
	[0.77 ; 1.1]	[0.88 ; 1.09]	[0.84 ; 1.12]	[0.92 ; 1.09]	[0.8 ; 1.15]	[0.89 ; 1.11]	[0.86 ; 1.1]	[0.91 ; 1.11]	[0.7 ; 1.17]	[0.82 ; 1.14]
Control Function (CF)	-0.671		-0.758		-0.702		-0.689		-0.743	
	[-0.82 ; -0.59]		[-0.87 ; -0.64]		[-0.83 ; -0.57]		[-0.78 ; -0.56]		[-0.87 ; -0.59]	
Intercept	-0.001	-0.002	-0.004	0.000	-0.006	-0.002	0.010	0.006	-0.005	-0.001
	[-0.02 ; 0.04]	[-0.02 ; 0.03]	[-0.02 ; 0.03]	[-0.02 ; 0.02]	[-0.02 ; 0.04]	[-0.02 ; 0.03]	[-0.03 ; 0.03]	[-0.03 ; 0.02]	[-0.03 ; 0.05]	[-0.03 ; 0.04]
Socioeconomic Index (SES)	-0.071	-0.019	-0.079	-0.052	-0.060	-0.007	-0.070	-0.028	-0.045	0.022
	[-0.17 ; 0.05]	[-0.11 ; 0.07]	[-0.18 ; 0.02]	[-0.16 ; 0.03]	[-0.16 ; 0.03]	[-0.11 ; 0.08]	[-0.14 ; 0.05]	[-0.11 ; 0.08]	[-0.16 ; 0.07]	[-0.1 ; 0.12]
Male	-0.091	-0.031	-0.118	-0.099	-0.096	-0.043	-0.102	-0.034	-0.005	0.044
	[-0.16 ; -0.02]	[-0.1 ; 0.04]	[-0.19 ; -0.05]	[-0.16 ; -0.03]	[-0.17 ; -0.03]	[-0.11 ; 0.03]	[-0.16 ; -0.02]	[-0.1 ; 0.05]	[-0.09 ; 0.09]	[-0.04 ; 0.14]
White	0.066	0.037	0.079	0.052	0.081	0.063	0.098	0.053	0.031	-0.003
	[0.01 ; 0.15]	[-0.03 ; 0.12]	[0.01 ; 0.16]	[-0.02 ; 0.12]	[0.02 ; 0.15]	[-0.01 ; 0.13]	[0 ; 0.15]	[-0.04 ; 0.11]	[-0.07 ; 0.13]	[-0.09 ; 0.08]
Number of siblings	-0.047	-0.064	-0.072	-0.099	-0.067	-0.079	-0.047	-0.060	-0.045	-0.061
	[-0.14 ; 0.02]	[-0.14 ; 0.01]	[-0.15 ; 0.02]	[-0.17 ; -0.02]	[-0.14 ; 0.02]	[-0.15 ; 0.01]	[-0.14 ; 0.02]	[-0.15 ; 0.01]	[-0.15 ; 0.06]	[-0.16 ; 0.03]
Parents' highest education	0.090	0.144	0.064	0.125	0.079	0.154	0.078	0.137	0.029	0.103
	[-0.04 ; 0.18]	[0.04 ; 0.23]	[-0.05 ; 0.17]	[0.03 ; 0.22]	[-0.03 ; 0.18]	[0.06 ; 0.25]	[-0.03 ; 0.17]	[0.04 ; 0.23]	[-0.09 ; 0.14]	[0 ; 0.21]

Note: each column estimates corresponds to a different synthetic dataset. Estimates with and without CF run in the same data. CF is the residual of investment equation that does not include age. 95% confidence intervals obtained using 500 bootstrap replications in square brackets.

B.1 Without Literacy (2008) as a measure for cognitive skills

Considering the years between 2008 and 2012 as being a single period is a strong assumption of our model. We assume that cognition remained stable in the period, and is assessed by three measures collected in different years, 2008 and 2012. If this assumption fails and cognitive skills distribution changed dramatically within the period, considering Literacy (2008), Portuguese and Math scores (2012) altogether could be misleading.

If we set aside Literacy score as a measure for cognitive skills, our measurement system is still identified [1]. Excluding Literacy from all steps of the estimation [2] is a way to verify the robustness of the estimated cognitive factor, but at the cost of loosing precious information from our database.

We consider Portuguese and Math scores only and normalize the factor loading in terms of Portuguese proficiency score. The distribution of $ln(\theta^C)$ without Literacy is pretty similar to when we include Literacy as a measure.

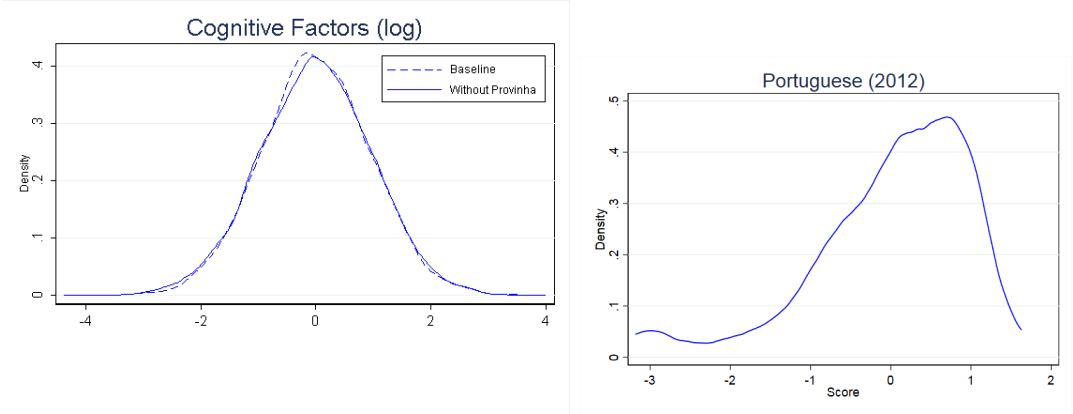
The factor loadings are pretty similar to our baseline specification: Portuguese proficiency remains more informative than Math, meanwhile the non-cognitive measures are pretty much the same.

The variables of the synthetic database are estimated altogether, thus running the algorithm without Literacy as a measure for cognition leads to small differences all variables, not only the cognitive factor. Still, we get very similar coefficients both in investment and CES equations.

After all, we opt to consider Literacy (Provinha) as a measure in our baseline specification in order to consider all available information from our original data, although disconsidering it would not significatively affect our estimates.

[1] Since we use a dedicated measurement system and assume that the measurement errors are independent across factors, and also between the measures of the same factor, being $\Sigma\varepsilon$ a diagonal matrix (CUNHA; HECKMAN; SCHENNACH, 2010)

[2] If we discard it as a measure for cognitive skill, we cannot use it as a control variable either because of its measurement error.



(a) Latent factors from synthetic dataset (10,000 obs)

(b) Normalized measure (1044 obs)

Figure 11: Cognitive factor distribution, estimated with and without Literacy as a measure

Table 21: Factor Loadings - without Literacy

		Baseline		Without Literacy	
Latent Factor	Measures	Loading	Signal to Noise	Loading	Signal to Noise
Cognitive skills	Portuguese Score	1.00	89.5%	1.00	97.2%
	Math Score	0.86	68.7%	0.76	55.3%
	Literacy Score	0.57	31.1%		
Non-cognitive skills	Conscientiousness	1.00	60.6%	1.00	60.0%
	Neurocitism	0.90	49.4%	0.91	50.0%
	Openness	1.00	59.0%	1.00	58.5%
	Agreeableness	1.04	62.6%	1.06	63.6%
	Extraversion	0.91	48.9%	0.91	48.4%

Table 22: Investment Equation Coefficients - without Literacy

Investment equation coefficients	Baseline	Without Literacy
Instrument (smallest distance to ECC)	-0.048	-0.047
	[-0.1 ; 0.01]	[-0.11 ; 0.01]
Socioeconomic Index (SES)	0.116	0.078
	[0 ; 0.18]	[-0.01 ; 0.16]
Male	-0.028	-0.025
	[-0.08 ; 0.04]	[-0.08 ; 0.04]
White	0.023	-0.014
	[-0.05 ; 0.07]	[-0.06 ; 0.07]
Number of siblings	-0.096	-0.085
	[-0.14 ; 0]	[-0.15 ; -0.01]
Parents' highest education	0.148	0.176
	[0.09 ; 0.26]	[0.09 ; 0.27]
Parents married in early childhood	-0.047	-0.042
	[-0.11 ; 0.03]	[-0.11 ; 0.03]
Seriously ill in early childhood	-0.019	-0.025
	[-0.08 ; 0.04]	[-0.09 ; 0.05]
Age	-0.052	-0.042
	[-0.1 ; 0.03]	[-0.11 ; 0.01]

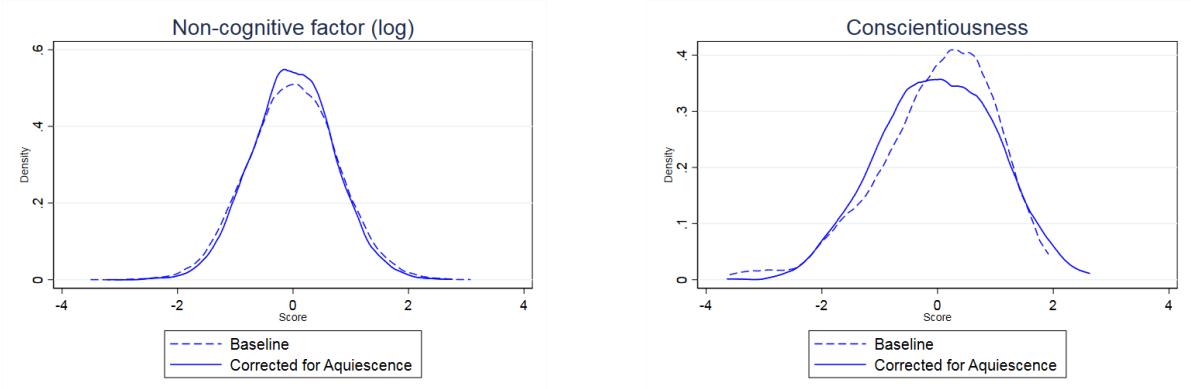
Note: different synthetic datasets generated for baseline and without Literacy as a measure for cognition. Estimates from columns (a) and (b) are run in the same dataset.
95% confidence intervals obtained using 500 bootstrap replications in square brackets.

Table 23: CES Production Function - without Literacy

CES coefficients	Baseline		Without Literacy	
	CF	no CF	CF	no CF
Investment (γ)	0.715	0.289	0.719	0.286
	[0.64 ; 0.84]	[0.24 ; 0.37]	[0.62 ; 0.81]	[0.22 ; 0.36]
Non-cognitive skills ($1-\gamma$)	0.285	0.711	0.281	0.714
	[0.16 ; 0.36]	[0.63 ; 0.76]	[0.19 ; 0.38]	[0.64 ; 0.78]
Complementarity (ρ)	0.013	0.009	-0.045	-0.066
	[-0.29 ; 0.09]	[-0.14 ; 0.08]	[-0.16 ; 0.09]	[-0.11 ; 0.1]
Elasticity Substitution (σ)	1.013	1.009	0.957	0.938
	[0.77 ; 1.1]	[0.88 ; 1.09]	[0.86 ; 1.1]	[0.9 ; 1.11]
Control Function (CF)	-0.668		-0.685	
	[-0.82 ; -0.59]		[-0.78 ; -0.55]	
Intercept	-0.001	-0.002	0.010	0.006
	[-0.02 ; 0.04]	[-0.02 ; 0.03]	[-0.03 ; 0.03]	[-0.03 ; 0.02]
Socioeconomic Index (SES)	-0.069	-0.015	-0.067	-0.023
	[-0.16 ; 0.05]	[-0.11 ; 0.08]	[-0.14 ; 0.05]	[-0.11 ; 0.09]
Male	-0.088	-0.026	-0.098	-0.027
	[-0.16 ; -0.02]	[-0.09 ; 0.05]	[-0.16 ; -0.02]	[-0.09 ; 0.06]
White	0.065	0.035	0.096	0.051
	[0.01 ; 0.15]	[-0.03 ; 0.11]	[0 ; 0.15]	[-0.04 ; 0.11]
Age	-0.028	-0.055	-0.041	-0.067
	[-0.13 ; 0.04]	[-0.15 ; 0]	[-0.13 ; 0.03]	[-0.14 ; 0]
Number of siblings	-0.045	-0.060	-0.043	-0.053
	[-0.13 ; 0.02]	[-0.14 ; 0.02]	[-0.13 ; 0.03]	[-0.14 ; 0.02]
Parents' highest education	0.085	0.134	0.071	0.125
	[-0.04 ; 0.17]	[0.03 ; 0.21]	[-0.04 ; 0.16]	[0.02 ; 0.22]

Note: different synthetic datasets generated for baseline specification and for specification without Provinha as a measure for cognition. Estimates with and without CF run in the same data.
95% confidence intervals obtained using 500 bootstrap replications in square brackets.

B.2 BFI corrected for acquiescence bias



(a) Latent factors from synthetic dataset (10,000 obs)

(b) Normalized measures (1044 obs)

Figure 12: Distributions of non-cognitive latent factors and normalized measures, with and without correction for acquiescence bias

Table 24: Factor Loadings - baseline *versus* BFI constructs corrected for acquiescence

		Baseline		Corrected for Aquiescence	
Latent Factor	Measures	Loading	Signal to Noise	Loading	Signal to Noise
Cognitive skills	Portuguese Score	1.00	89.5%	1.00	88.7%
	Math Score	0.86	68.7%	0.88	70.6%
	Literacy Score	0.57	31.1%	0.56	30.1%
Non-cognitive skills	Conscientiousness	1.00	60.6%	1.00	52.3%
	Neurocitism	0.90	49.4%	0.93	44.3%
	Openness	1.00	59.0%	1.13	65.2%
	Agreeableness	1.04	62.6%	1.11	59.9%
	Extraversion	0.91	48.9%	0.92	42.4%

Table 25: Investment Equation Coefficients - baseline *versus* BFI constructs corrected for acquiescence

Investment equation coefficients	Baseline	Corrected for Aquiescence
Instrument (smallest distance to ECC)	-0.048	-0.043
	[-0.1 ; 0.01]	[-0.1 ; 0.02]
Socioeconomic Index (SES)	0.116	0.098
	[0 ; 0.18]	[0.01 ; 0.17]
Male	-0.028	-0.033
	[-0.08 ; 0.04]	[-0.08 ; 0.04]
White	0.023	0.018
	[-0.05 ; 0.07]	[-0.05 ; 0.07]
Number of siblings	-0.096	-0.075
	[-0.14 ; 0]	[-0.15 ; -0.01]
Parents' highest education	0.148	0.170
	[0.09 ; 0.26]	[0.09 ; 0.26]
Parents married in early childhood	-0.047	-0.040
	[-0.11 ; 0.03]	[-0.1 ; 0.02]
Seriously ill in early childhood	-0.019	-0.035
	[-0.08 ; 0.04]	[-0.08 ; 0.05]
Age	-0.052	-0.052
	[-0.1 ; 0.03]	[-0.1 ; 0.03]

Note: different synthetic datasets generated for baseline and using BFI corrected for individual acquiescence bias. 95% confidence intervals obtained using 500 bootstrap replications in square brackets.

Table 26: CES Production Function - baseline *versus* BFI constructs corrected for acquiescence

CES coefficients	Baseline		Corrected for Aquiescence	
	CF	no CF	CF	no CF
Investment (γ)	0.715	0.289	0.750	0.294
	[0.64 ; 0.84]	[0.24 ; 0.37]	[0.64 ; 0.87]	[0.22 ; 0.38]
Non-cognitive skills ($1-\gamma$)	0.285	0.711	0.250	0.706
	[0.16 ; 0.36]	[0.63 ; 0.76]	[0.13 ; 0.36]	[0.62 ; 0.78]
Complementarity (ρ)	0.013	0.009	0.029	0.019
	[-0.29 ; 0.09]	[-0.14 ; 0.08]	[-0.25 ; 0.13]	[-0.12 ; 0.1]
Elasticity Substitution (σ)	1.013	1.009	1.030	1.019
	[0.77 ; 1.1]	[0.88 ; 1.09]	[0.8 ; 1.14]	[0.89 ; 1.11]
Control Function (CF)	-0.668		-0.700	
	[-0.82 ; -0.59]		[-0.83 ; -0.57]	
Intercept	-0.001	-0.002	-0.006	-0.002
	[-0.02 ; 0.04]	[-0.02 ; 0.03]	[-0.02 ; 0.04]	[-0.02 ; 0.03]
Socioeconomic Index (SES)	-0.069	-0.015	-0.059	-0.005
	[-0.16 ; 0.05]	[-0.11 ; 0.08]	[-0.16 ; 0.04]	[-0.11 ; 0.09]
Male	-0.088	-0.026	-0.095	-0.040
	[-0.16 ; -0.02]	[-0.09 ; 0.05]	[-0.17 ; -0.02]	[-0.11 ; 0.03]
White	0.065	0.035	0.080	0.062
	[0.01 ; 0.15]	[-0.03 ; 0.11]	[0.02 ; 0.15]	[-0.01 ; 0.13]
Age	-0.028	-0.055	-0.017	-0.034
	[-0.13 ; 0.04]	[-0.15 ; 0]	[-0.12 ; 0.04]	[-0.13 ; 0.03]
Number of siblings	-0.045	-0.060	-0.065	-0.076
	[-0.13 ; 0.02]	[-0.14 ; 0.02]	[-0.14 ; 0.02]	[-0.15 ; 0.02]
Parents' highest education	0.085	0.134	0.076	0.148
	[-0.04 ; 0.17]	[0.03 ; 0.21]	[-0.04 ; 0.18]	[0.05 ; 0.25]

Note: different synthetic datasets generated for baseline and using BFI corrected for individual acquiescence bias. Estimates with and without CF run in the same data. 95% confidence intervals obtained using 500 bootstrap replications in square brackets.

B.3 Using alternative measures for non-cognitive factor

In this section we present results using source available in our dataset to assess socioemotional skills. Instead of the BFI questionnaire, we considered the answers of 34 items, for which children should assign 0 if the statement never applied to them, 1 if a few times or 2 if the statement reflected their behavior very often. Each item was designed to assess one specific Social Skill (SS), namely Cooperation, Assertiveness, Empathy and Self-control, that were extracted using Cronbach's Alpha, without any adjustment in the scores. Their distribution is shown in figure 13, and we normalize the resulting non-cognitive factor in units of Cooperation, instead of the Conscientiousness construct from the BFI.

The latent factor drawn is very similar to the baseline specification. Although the Big Five Personality Traits and the Social Skills measure different traits in different scales, it seems that both of them are assessing the same socioemotional latent factor. The same goes to the factor loadings, despite the signals of social skills' measures be a bit higher than the baseline specification.

There are minor changes in the investment equation and in the CES coefficients. The share of the socioemotional factor obtained from SS measures is slightly above the baseline specification, both with and without the control function.

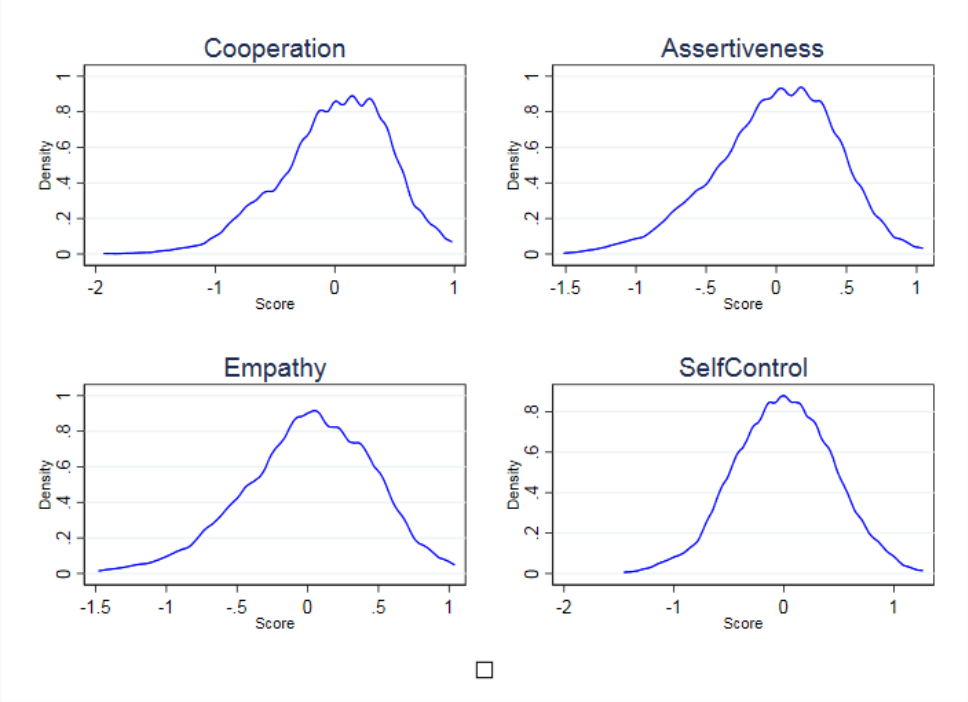
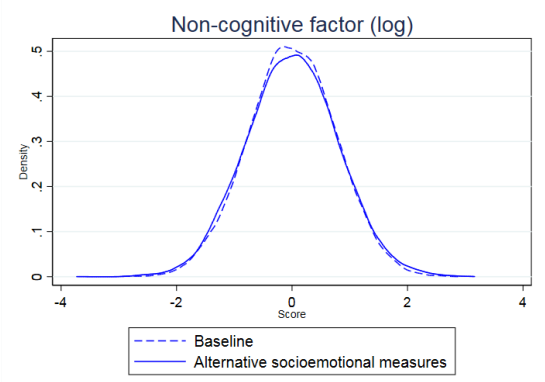
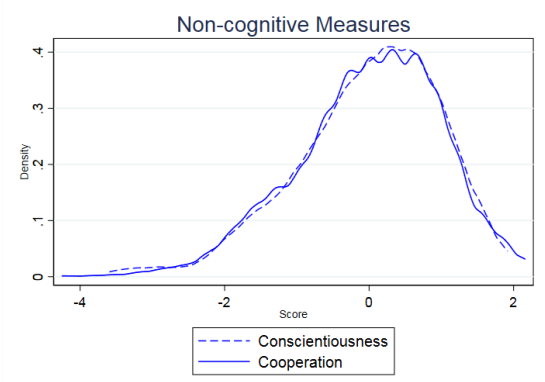


Figure 13: Distribution of Social Skills' Measures



(a) Latent factors from synthetic dataset (10,000 obs)



(b) Normalized measures (1044 obs)

Figure 14: Distributions of non-cognitive latent factors and normalized measures, with alternative measures (Social Skills)

Table 27: Factor Loadings - with alternative measures (Social Skills) for non-cognitive factor

Latent Factor	Measures	Loading	Signal to Noise
Cognitive skills	Portuguese Score	1.00	79.6%
	Math Score	0.93	72.4%
	Literacy Score	0.63	35.2%
Non-cognitive skills	Cooperation	1.00	63.5%
	Assertiveness	1.01	66.6%
	Empathy	0.92	51.5%
	Self-control	0.94	57.9%

Table 28: Investment Equation Coefficients - with alternative measures (Social Skills) for non-cognitive factor

Investment equation coefficients	Baseline	Social Skills
Instrument (smallest distance to ECC)	-0.048	-0.043
	[-0.1 ; 0.01]	[-0.11 ; 0.02]
Socioeconomic Index (SES)	0.116	0.099
	[0 ; 0.18]	[0.01 ; 0.17]
Male	-0.028	-0.032
	[-0.08 ; 0.04]	[-0.08 ; 0.04]
White	0.023	0.018
	[-0.05 ; 0.07]	[-0.05 ; 0.07]
Number of siblings	-0.096	-0.075
	[-0.14 ; 0]	[-0.15 ; -0.01]
Parents' highest education	0.148	0.170
	[0.09 ; 0.26]	[0.09 ; 0.26]
Parents married in early childhood	-0.047	-0.040
	[-0.11 ; 0.03]	[-0.1 ; 0.02]
Seriously ill in early childhood	-0.019	-0.035
	[-0.08 ; 0.04]	[-0.08 ; 0.04]
Age	-0.052	-0.052
	[-0.1 ; 0.03]	[-0.1 ; 0.03]

Note: estimates from each column generated from a different synthetic dataset. 95% confidence intervals obtained using 500 bootstrap replications in square brackets.

Table 29: CES Production Function - with alternative measures (Social Skills) for non-cognitive factor

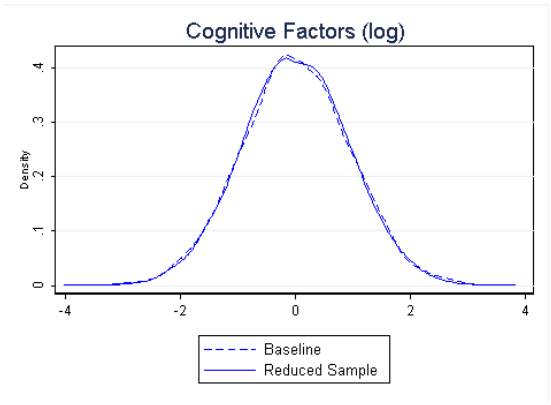
CES coefficients	Baseline		Social Skills	
	CF	no CF	CF	no CF
Investment (γ)	0.715	0.289	0.809	0.362
	[0.64 ; 0.84]	[0.24 ; 0.37]	[0.72 ; 0.9]	[0.3 ; 0.42]
Non-cognitive skills ($1-\gamma$)	0.285	0.711	0.191	0.638
	[0.16 ; 0.36]	[0.63 ; 0.76]	[0.1 ; 0.28]	[0.58 ; 0.7]
Complementarity (ρ)	0.013	0.009	0.017	-0.002
	[-0.29 ; 0.09]	[-0.14 ; 0.08]	[-0.19 ; 0.11]	[-0.09 ; 0.08]
Elasticity Substitution (σ)	1.013	1.009	1.017	0.998
	[0.77 ; 1.1]	[0.88 ; 1.09]	[0.84 ; 1.12]	[0.92 ; 1.09]
Control Function (CF)	-0.668		-0.758	
	[-0.82 ; -0.59]		[-0.87 ; -0.64]	
Intercept	-0.001	-0.002	-0.004	0.000
	[-0.02 ; 0.04]	[-0.02 ; 0.03]	[-0.02 ; 0.03]	[-0.02 ; 0.02]
Socioeconomic Index (SES)	-0.069	-0.015	-0.078	-0.052
	[-0.16 ; 0.05]	[-0.11 ; 0.08]	[-0.18 ; 0.02]	[-0.16 ; 0.03]
Male	-0.088	-0.026	-0.117	-0.098
	[-0.16 ; -0.02]	[-0.09 ; 0.05]	[-0.19 ; -0.04]	[-0.16 ; -0.02]
White	0.065	0.035	0.079	0.051
	[0.01 ; 0.15]	[-0.03 ; 0.11]	[0.01 ; 0.16]	[-0.02 ; 0.12]
Age	-0.028	-0.055	-0.010	-0.016
	[-0.13 ; 0.04]	[-0.15 ; 0]	[-0.11 ; 0.05]	[-0.12 ; 0.04]
Number of siblings	-0.045	-0.060	-0.071	-0.098
	[-0.13 ; 0.02]	[-0.14 ; 0.02]	[-0.14 ; 0.02]	[-0.17 ; -0.02]
Parents' highest education	0.085	0.134	0.062	0.123
	[-0.04 ; 0.17]	[0.03 ; 0.21]	[-0.06 ; 0.16]	[0.02 ; 0.22]

Note: estimates from each pair of columns generated from different synthetic datasets. Estimates with and without CF run in the same data. 95% confidence intervals obtained using 500 bootstrap replications in square brackets.

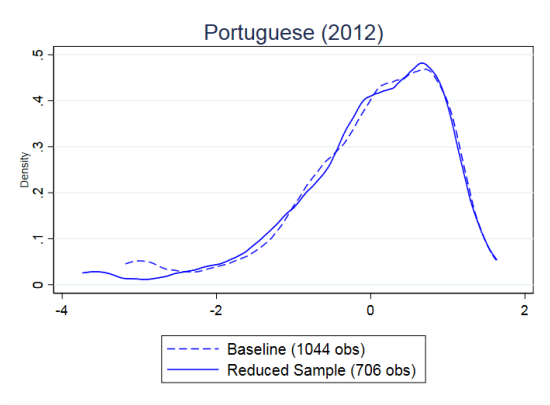
B.4 Using a subgroup of our sample (706 observations)

Since we only have long-term cognitive measures for a subgroup of our data, we build a synthetic dataset from this 706 observations and draw the distribution of the latent factors as in the baseline case. By doing so, we can have an idea of how the estimates of this subgroup compares to our full data.

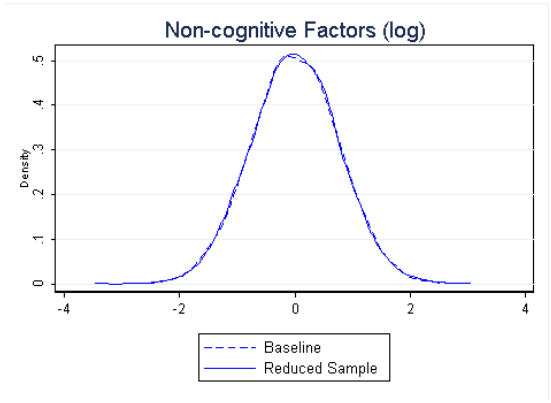
Results are reasonably stable comparing to the full sample estimates. The magnitude of the key coefficients do not change dramatically, although some change signal. As expected, even with 500 bootstraps, confidence intervals are wider in the reduced sample.



(a) Latent factors from synthetic dataset (10,000 obs)



(b) Normalized measures



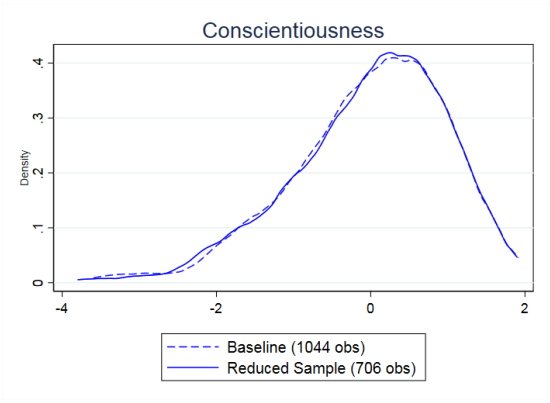


Figure 15: Distributions of latent factors and normalized measures, from baseline (full sample,1044 obs) and reduced sample (706 obs)

Table 30: Factor Loadings - baseline (1044 obs) *versus* reduced sample (706 obs)

		Baseline: Raw data=1044 obs		Raw data = 706 obs	
Latent Factor	Measures	Loading	Signal to Noise	Loading	Signal to Noise
Cognitive skills	Portuguese Score	1.00	89.5%	1.00	85.7%
	Math Score	0.86	68.7%	0.85	63.4%
	Literacy Score	0.57	31.1%	0.62	38.7%
Non-cognitive skills	Conscientiousness	1.00	60.6%	1.00	57.1%
	Neurocitism	0.90	49.4%	0.90	47.8%
	Openness	1.00	59.0%	1.03	60.0%
	Agreeableness	1.04	62.6%	1.06	63.4%
	Extraversion	0.91	48.9%	0.90	45.5%

Table 31: Investment Equation Coefficients - baseline *versus* reduced sample

Investment equation coefficients	Baseline:Raw data=1044 obs	Raw data = 706 obs
Instrument (smallest distance to ECC)	-0.048	-0.048
	[-0.1 ; 0.01]	[-0.12 ; 0.01]
Socioeconomic Index (SES)	0.116	0.078
	[0 ; 0.18]	[-0.04 ; 0.16]
Male	-0.028	-0.073
	[-0.08 ; 0.04]	[-0.14 ; 0.01]
White	0.023	0.020
	[-0.05 ; 0.07]	[-0.06 ; 0.09]
Number of siblings	-0.096	-0.073
	[-0.14 ; 0]	[-0.16 ; 0.01]
Parents' highest education	0.148	0.190
	[0.09 ; 0.26]	[0.09 ; 0.3]
Parents married in early childhood	-0.047	-0.022
	[-0.11 ; 0.03]	[-0.1 ; 0.05]
Seriously ill in early childhood	-0.019	-0.043
	[-0.08 ; 0.04]	[-0.11 ; 0.05]
Age	-0.052	-0.067
	[-0.1 ; 0.03]	[-0.14 ; 0.03]

Note: different synthetic datasets with 10,000 observations generated for full sample (1044 obs) and for the subgroup (706 observations). 95% confidence intervals obtained using 500 bootstrap replications in square brackets.

Table 32: CES Production Function - baseline *versus* reduced sample

CES coefficients	Baseline: Raw data=1044 obs		Raw data = 706 obs	
	CF	no CF	CF	no CF
Investment (γ)	0.715	0.289	0.761	0.297
	[0.64 ; 0.84]	[0.24 ; 0.37]	[0.62 ; 0.88]	[0.2 ; 0.4]
Non-cognitive skills ($1-\gamma$)	0.285	0.711	0.239	0.703
	[0.16 ; 0.36]	[0.63 ; 0.76]	[0.12 ; 0.38]	[0.6 ; 0.8]
Complementarity (ρ)	0.013	0.009	0.018	0.002
	[-0.29 ; 0.09]	[-0.14 ; 0.08]	[-0.42 ; 0.15]	[-0.21 ; 0.12]
Elasticity Substitution (σ)	1.013	1.009	1.019	1.002
	[0.77 ; 1.1]	[0.88 ; 1.09]	[0.7 ; 1.17]	[0.83 ; 1.14]
Control Function (CF)	-0.668		-0.738	
	[-0.82 ; -0.59]		[-0.87 ; -0.58]	
Intercept	-0.001	-0.002	-0.005	-0.001
	[-0.02 ; 0.04]	[-0.02 ; 0.03]	[-0.03 ; 0.05]	[-0.03 ; 0.04]
Socioeconomic Index (SES)	-0.069	-0.015	-0.042	0.029
	[-0.16 ; 0.05]	[-0.11 ; 0.08]	[-0.16 ; 0.08]	[-0.09 ; 0.14]
Male	-0.088	-0.026	-0.004	0.048
	[-0.16 ; -0.02]	[-0.09 ; 0.05]	[-0.09 ; 0.09]	[-0.04 ; 0.15]
White	0.065	0.035	0.030	-0.008
	[0.01 ; 0.15]	[-0.03 ; 0.11]	[-0.07 ; 0.13]	[-0.1 ; 0.08]
Age	-0.028	-0.055	-0.024	-0.080
	[-0.13 ; 0.04]	[-0.15 ; 0]	[-0.14 ; 0.06]	[-0.18 ; -0.01]
Number of siblings	-0.045	-0.060	-0.044	-0.058
	[-0.13 ; 0.02]	[-0.14 ; 0.02]	[-0.15 ; 0.06]	[-0.16 ; 0.03]
Parents' highest education	0.085	0.134	0.027	0.092
	[-0.04 ; 0.17]	[0.03 ; 0.21]	[-0.09 ; 0.14]	[-0.01 ; 0.2]

Note: different synthetic datasets with 10,000 observations generated for full sample (1044 obs) and for the subgroup (706 observations). Estimates with and without CF run in the same data. 95% confidence intervals obtained using 500 bootstrap replications in square brackets.

C – Approximation of the impact of investment in long-term cognition

Our model comprises the following production functions (we omit subscript i for simplicity):

$$\theta_2^C = A[\gamma(I_1)^\rho + (1-\gamma)(\theta^N)^\rho]^{\frac{1}{\rho}} e^{\eta_1} \tag{C.1}$$

$$\theta_3^C = A[\delta(\theta_2^C)^\varrho + (1-\delta)(\theta^N)^\varrho]^{\frac{1}{\varrho}} e^{\eta_2} \tag{C.2}$$

We can decompose θ_3^C in the part that is due to inputs θ_3^C and θ_3^N:

$$\theta_3^C = \frac{\partial \theta_3^C}{\partial \theta_2^C}\theta_2^C + \frac{\partial \theta_3^C}{\partial \theta^N}\theta^N \tag{C.3}$$

The above equation is true since

$$\begin{aligned}\frac{\partial \theta_3^C}{\partial \theta_2^C}\theta_2^C &= \frac{1}{\varrho} A[\delta(\theta_2^C)^\varrho + (1-\delta)(\theta^N)^\varrho]^{(\frac{1}{\varrho}-1)} e^{\eta_2}\ \varrho\delta(\theta_2^C)^{\varrho-1}\theta_2^C \\ &= \theta_3^C \frac{\delta(\theta_2^C)^\varrho}{[\delta(\theta_2^C)^\varrho + (1-\delta)(\theta^N)^\varrho]} \qquad \text{(F.4)} \\ \frac{\partial \theta_3^C}{\partial \theta^N}\theta^N &= \frac{1}{\varrho} A[\delta(\theta_2^C)^\varrho + (1-\delta)(\theta^N)^\varrho]^{(\frac{1}{\varrho}-1)} e^{\eta_2} \varrho(1-\delta)(\theta_2^N)^{\varrho-1}\theta_2^N \\ &= \theta_3^C \frac{(1-\delta)(\theta^N)^\varrho}{[\delta(\theta_2^C)^\varrho + (1-\delta)(\theta^N)^\varrho]} \qquad \text{(F.5)}\end{aligned}$$

Then, substituting (F.4) and (F.5) in (F.3) gives

$$\theta_3^C = \theta_3^C \frac{\delta(\theta_2^C)^\varrho + (1-\delta)(\theta^N)^\varrho}{[\delta(\theta_2^C)^\varrho + (1-\delta)(\theta^N)^\varrho]} = \theta_3^C$$

Analogously to (F.3), we can also decompose θ_2^C into the effects of I_1 and θ^N:

$$\theta_2^C = \frac{\partial \theta_2^C}{\partial I_1} I_1 + \frac{\partial \theta_2^C}{\partial \theta^N}\theta^N \tag{F.6}$$

Putting (F.6) in (F.3):

$$\theta_3^C = \frac{\partial \theta_3^C}{\partial \theta_2^C}\left(\frac{\partial \theta_2^C}{\partial I_1} I_1 + \frac{\partial \theta_2^C}{\partial \theta^N}\theta^N\right) + \frac{\partial \theta_3^C}{\partial \theta^N}\theta^N \tag{F.7}$$

By dividing both sides of equation (F.7), we get that the fraction of θ_3^C that is due to the investment equals to

$$share_{I_1} = \frac{\partial \theta_3^C}{\partial \theta_2^C}\frac{\partial \theta_2^C}{\partial I_1}\frac{I_1}{\theta_3^C} \tag{F.8}$$

Intuitively, the impact of investment in the long-term cognition $\left(\frac{\partial \theta_3^C}{\partial \theta_2^C}\frac{\partial \theta_2^C}{\partial I_1}\right)$ evaluated at a given level of investment is the magnitude of the cumulative effect of the investment. To know how much it represents of the level of the long-term cognition, we divide by θ_3^C. This is what we get in (F.8).

Finally, if we calculate the partial derivatives and simplify (F.8), we get

$$share_{I_1} = \frac{\delta(\theta_2^C)^{\varrho}}{[\delta(\theta_2^C)^{\varrho} + (1-\delta)(\theta^N)^{\varrho}]}\ \frac{\gamma(I_1)^{\rho}}{[\gamma(I_1)^{\rho} + (1-\gamma)(\theta^N)^{\rho}]} \tag{F.9}$$

D – Further descriptive statistics for the raw database

Table 33: Descriptive statistics by school entry age: daycare *versus* elementary school

	Daycare (0-3)		**Elementary School (6+)**		**Difference**
Observations	130		130		
Descriptive Statistics	Mean	SD	Mean	SD	
Age	7.555	0.575	7.723	0.671	0.168***
Male	0.448	0.498	0.508	0.502	0.06
White	0.440	0.497	0.338	0.475	-0.101**
Black	0.071	0.258	0.138	0.347	0.067**
Mother Education (highest level achieved)					
None	0.112	0.316	0.328	0.471	0.215***
Elementary School	0.330	0.471	0.431	0.497	0.101*
Middle School	0.202	0.402	0.164	0.372	-0.04
High school	0.277	0.448	0.078	0.269	-0.200***
College	0.078	0.268	0.000	0.000	-0.078***
Father Education (highest level achieved)					
None	0.112	0.316	0.291	0.457	0.179***
Elementary School	0.316	0.466	0.417	0.496	0.101*
Middle School	0.211	0.409	0.165	0.373	-0.05
High school	0.275	0.447	0.126	0.334	-0.149***
College	0.086	0.281	0.000	0.000	-0.086***
Household configuration					
Lives with both parents	0.636	0.482	0.591	0.494	-0.05
Number of people in the household	3.224	1.297	3.868	1.378	0.644***
Number of siblings	1.945	1.957	2.900	2.372	0.955***
That attended daycare	0.431	0.896	0.562	1.220	0.13
That attended kindergarten	0.555	1.004	0.785	1.175	0.230**
Has a dictionary at home	0.69	0.46	0.57	0.50	-0.122**
A shelf with 20+ books	0.29	0.45	0.17	0.38	-0.119**
Has children's books	0.79	0.41	0.65	0.48	-0.138***
Attended public school in 2008	0.86	0.35	0.98	0.15	0.117***
Dimensions of income					
Bolsa Familia beneficiary	0.182	0.386	0.391	0.490	0.209***
Other social benefit	0.081	0.274	0.148	0.357	0.067**
SES Index (calculated from data)	12.456	5.228	8.901	3.743	-3.555***

Table 34: Descriptive statistics by school entry age: kindergarten *versus* elementary school

	Kindergarten (4-5)		**Elementary School (6+)**		**Difference**
Observations	550		130		
Descriptive Statistics	Mean	SD	Mean	SD	
Age	7.585	0.612	7.723	0.671	0.138**
Male	0.514	0.500	0.508	0.502	-0.01
White	0.436	0.496	0.338	0.475	-0.098**
Black	0.080	0.272	0.138	0.347	0.058**
Mother Education (highest level achieved)					
None	0.089	0.284	0.328	0.471	0.239***
Elementary School	0.420	0.494	0.431	0.497	0.01
Middle School	0.230	0.421	0.164	0.372	-0.07
High school	0.232	0.422	0.078	0.269	-0.154***
College	0.030	0.171	0.000	0.000	-0.030*
Father Education (highest level achieved)					
None	0.128	0.334	0.291	0.457	0.163***
Elementary School	0.380	0.486	0.417	0.496	0.04
Middle School	0.234	0.424	0.165	0.373	-0.07
High school	0.230	0.421	0.126	0.334	-0.104**
College	0.028	0.164	0.000	0.000	-0.028*
Household configuration					
Lives with both parents	0.701	0.458	0.591	0.494	-0.110**
Number of people in the household	3.431	1.274	3.868	1.378	0.437***
Number of siblings	2.000	1.892	2.900	2.372	0.900***
That attended daycare	0.324	0.757	0.562	1.220	0.238***
That attended kindergarten	0.540	0.951	0.785	1.175	0.245**
Has a dictionary at home	0.71	0.45	0.57	0.50	-0.149***
A shelf with 20+ books	0.29	0.45	0.17	0.38	-0.116***
Has children's books	0.77	0.42	0.65	0.48	-0.124***
Attended public school in 2008	0.92	0.27	0.98	0.15	0.055**
Dimensions of income					
Bolsa Familia beneficiary	0.203	0.402	0.391	0.490	0.189***
Other social benefit	0.073	0.260	0.148	0.357	0.076***
SES Index (calculated from data)	11.684	4.369	8.901	3.743	-2.783***

Table 35: Cognitive and socioemotional measures differences according to school entry age: daycare *versus* elementary school

Average and distribution differences		Average Scores Homecare	Daycare	Difference	(K-S)p-value
Cognitive measures	Literacy	20.18	21.21	-1.032***	0.000
	Portuguese Score	60.52	68.35	-7.822***	0.000
	Math Score	52.28	61.19	-8.913***	0.000
Non-cognitive measures	Conscientiousness	-0.06	0.02	-0.08	0.220
	Neurocitism	-0.06	-0.01	-0.05	0.236
	Openness	-0.04	0.01	-0.05	0.166
	Agreeableness	-0.09	-0.01	-0.08	0.299
	Extraversion	-0.06	-0.01	-0.06	0.149

*Note:*** p<0.01, ** p<0.05, * p<0.10. Kolmogorov-Smirnov (K-S) null is of equal distributions.*

Table 36: Cognitive and socioemotional measures differences according to school entry age: kindergarten *versus* elementary school

Average and distribution differences		Average		Difference	(K-S)p-value
		Homecare	Kindergarten		
Cognitive measures	Literacy	20.18	20.91	-0.732**	0.001
	Portuguese Score	60.52	66.07	-5.543***	0.003
	Math Score	52.28	60.43	-8.152***	0.000
Non-cognitive measures	Conscientiousness	-0.06	0.00	-0.05	0.155
	Neurocitism	-0.06	0.01	-0.07	0.402
	Openness	-0.04	0.00	-0.04	0.580
	Agreeableness	-0.09	0.03	-0.112***	0.037
	Extraversion	-0.06	0.01	-0.076*	0.071

*Note:*** p<0.01, ** p<0.05, * p<0.10. Kolmogorov-Smirnov (K-S) null is of equal distributions.*

D.1 Reduced sample (706 observations)

Table 37: Descriptive statistics of full sample (1044 obs) *versus* reduced sample (706 obs)

	Reduced Sample		**Full Sample**		**Difference**
Observations	706		1044		
Descriptive Statistics	Mean	SD	Mean	SD	
Age	7.531	0.553	7.592	0.608	0.061**
Male	0.455	0.498	0.490	0.500	0.04
White	0.431	0.496	0.425	0.495	-0.01
Black	0.078	0.268	0.084	0.278	0.01
Mother Education (highest level achieved)					
None	0.104	0.306	0.125	0.331	0.02
Elementary School	0.371	0.484	0.392	0.488	0.02
Middle School	0.226	0.418	0.213	0.410	-0.01
High school	0.253	0.435	0.228	0.420	-0.02
College	0.046	0.210	0.042	0.202	0.00
Father Education (highest level achieved)					
None	0.118	0.322	0.141	0.348	0.02
Elementary School	0.334	0.472	0.363	0.481	0.03
Middle School	0.234	0.424	0.219	0.414	-0.02
High school	0.277	0.448	0.234	0.423	-0.043*
College	0.037	0.189	0.044	0.206	0.01
Household configuratoin					
Lives with both parents	0.682	0.466	0.665	0.472	-0.02
Number of people in the household	3.285	1.248	3.416	1.308	0.131**
Number of siblings	1.864	1.784	2.093	2.002	0.229**
That attended daycare	0.380	0.869	0.391	0.878	0.01
That attended kindergarten	0.557	0.998	0.576	1.002	0.02
Has a dictionary at home	0.70	0.46	0.69	0.46	-0.02
A shelf with 20+ books	0.28	0.45	0.27	0.45	-0.01
Has children's books	0.78	0.41	0.76	0.43	-0.02
Attended public school in 2008	0.90	0.30	0.91	0.29	0.00
Dimensions of income					
Bolsa Familia beneficiary	0.182	0.386	0.218	0.413	0.037*
Other social benefit	0.084	0.278	0.085	0.279	0.00
SES Index (calculated from data)	12.018	4.539	11.598	4.724	-0.420*

In this section we report how the sample of students that were already found in 2017 compares to our full sample. We observe no significant differences with respect to household characteristics.

The students that were already mapped in 2017 had performed better in Math and Portuguese exams than the full sample - indeed, we reject the null hypothesis that the full and reduced sample distributions are equal using Kolmogorov-Smirnov (K-S) test. Although the average of Literacy Score (figure 16) was higher in the reduced sample, we cannot reject that the two groups Literacy Scores belong to the same distribution. The reduced and full sample do not differ with respect to BFI (figure 17).

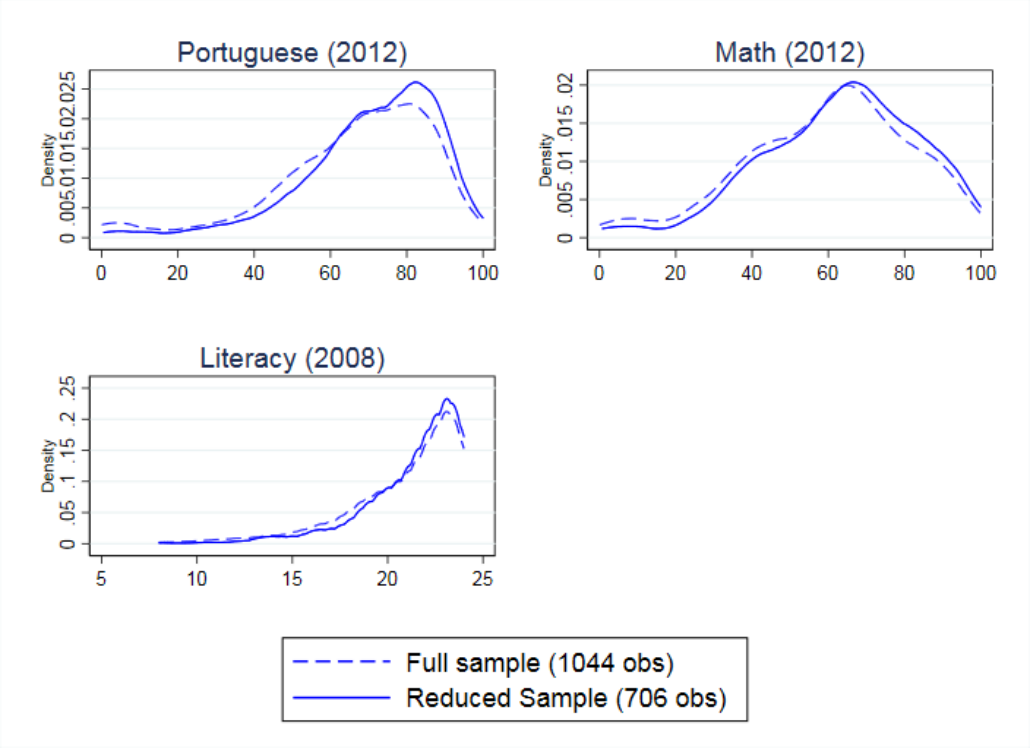


Figure 16: Portuguese and Math Proficiency Scores (2012) and Literacy Scores (Provinha 2008) distributions, full sample (1044 obs) *versus* reduced sample (706 obs)

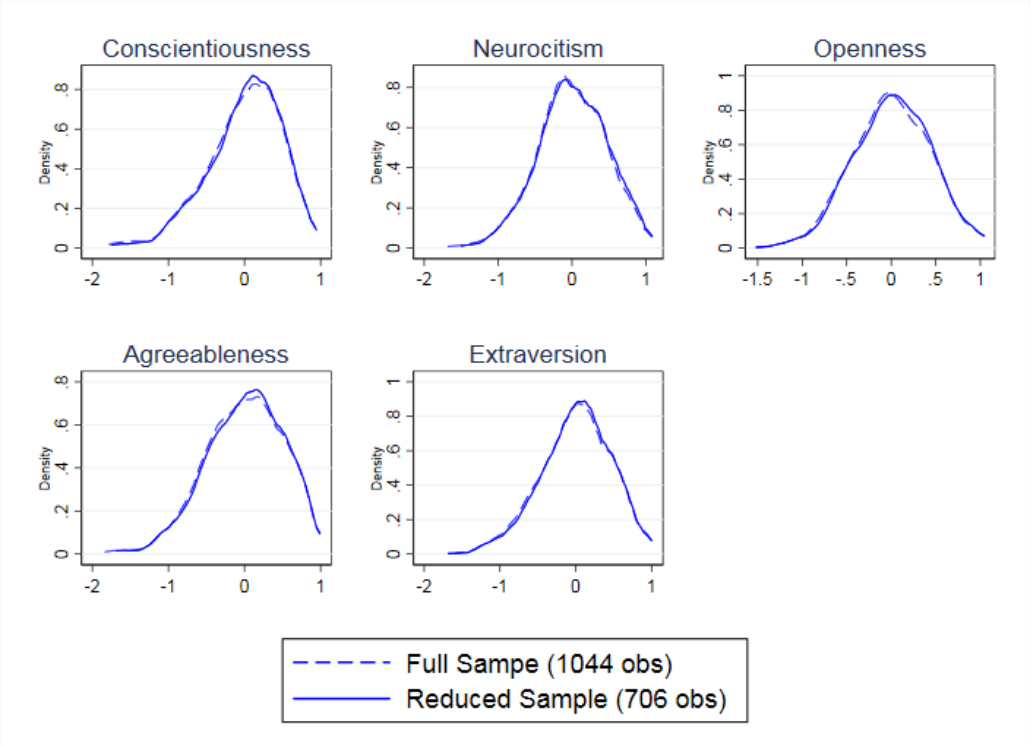


Figure 17: Big Five Personality Traits distribution, , full sample (1044 obs) *versus* reduced sample (706 obs)

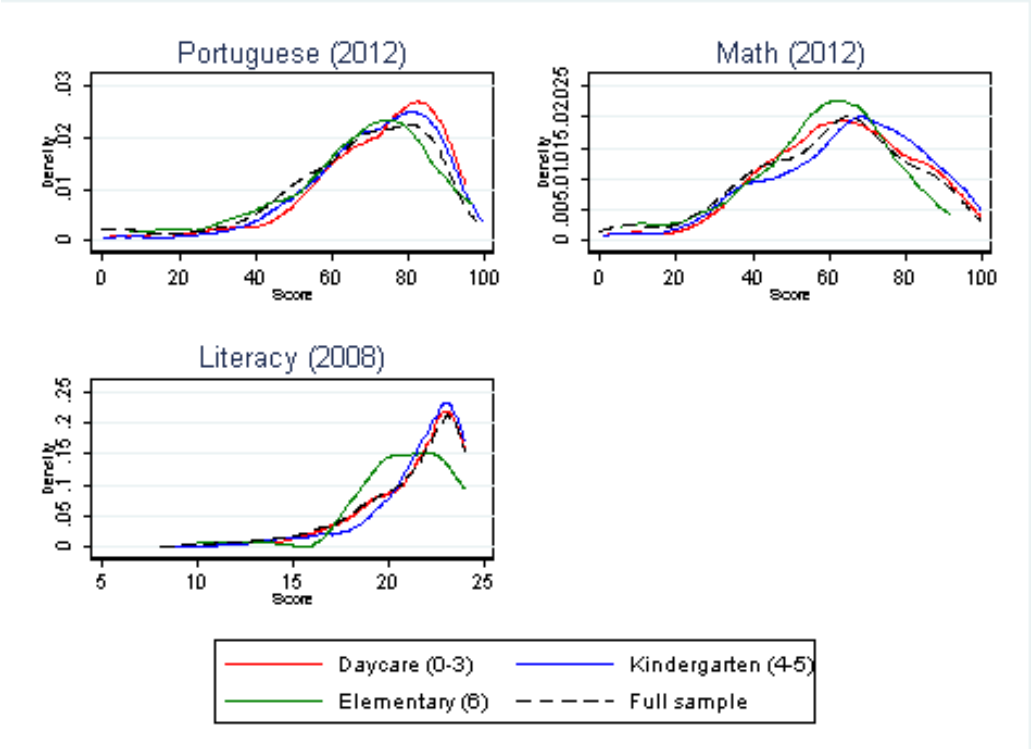


Figure 18: Portuguese and Math Proficiency Scores (2012) and Literacy Scores (Provinha 2008) distributions, by school entry age - reduced sample (706 obs)

If we take a closer look to cognitive and non-cognitive measures by school entry age (figure 18), as we did to the full sample, we still get children that only started school at mandatory ages (green line) performing below preschool attendees in Portuguese and Literacy but, surprisingly, would had been better off in Math, when comparing to all 706 children. For the case of the BFI, we still have not so clear picture of how preschool attendees and non-attendees would differ with respect socioemocional measures (figure 19).

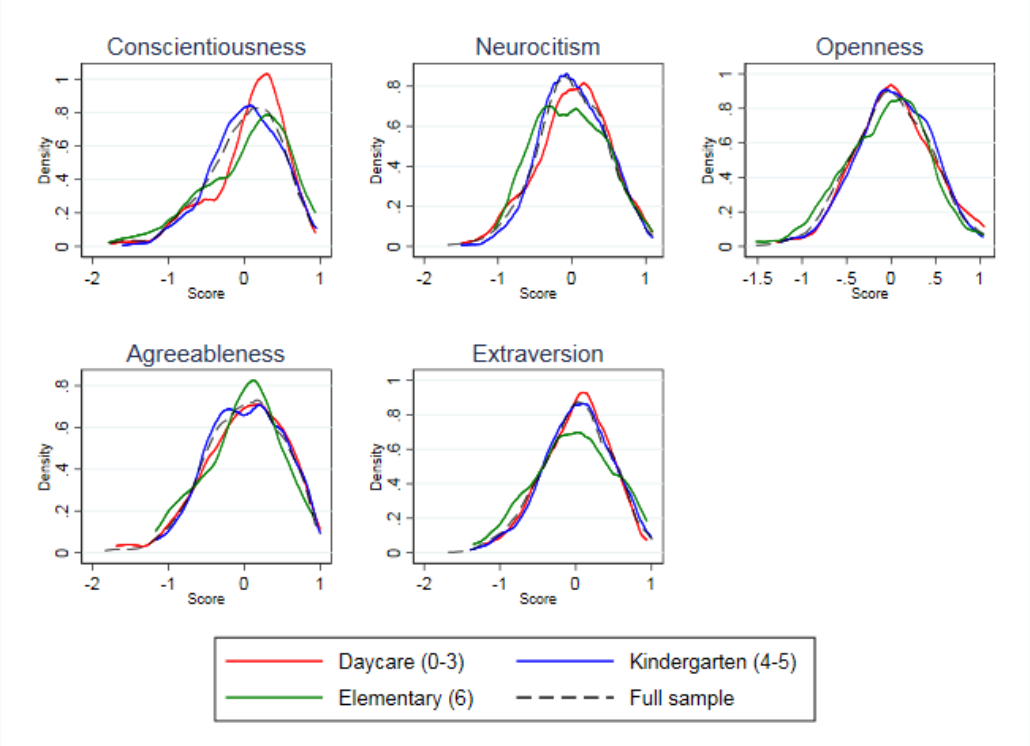


Figure 19: Big Five Personality Traits distribution, by school entry age, by school entry age - reduced sample (706 obs)

E – Instrument for preschool attendance

E.1 Estimating the area where students lived

We build the estimated area where students would live form questions answered by parents in 2012 questionnaire. We anchor the household estimated location to the address of the elementary schools attended in 2008, since it is the first reference of location we have for all the children in our sample. Moreover, the majority of the children in our sample go to public schools and have to change school in middle school[1]. This means that the school attended in 2012 might not be as convenient in terms of location as daycare/kindergarten or elementary school [2].

In 2012 questionnaire, parents reported how many minutes the child took to school during daycare, kindergarten, elementary school ($1^{st} - 5^{th}$ grade) and middle school ($6^{th} - 9^{th}$ grade). Parents also reported the main mean of transport used. To obtain the distance (in meters) that each child lived from school, we combine the reported distance (in minutes)[3] to the elementary school with the average speed reached by the corresponding mean of transport. This last information was obtained from Google Maps Distance Matrix API [4]. Building a function in Excel VBA, we stored the distance and duration of routes from the 20 elementary schools in our sample to the 27 schools that offered daycare and/or kindergarten [5].

Table 38 displays how the 2,713 recommended routes collected are distributed by mean of transportation and by the range of duration (in minutes). All these routes were collected during weekdays in business hours. Since Sertãozinho is relatively small, no distance between two addresses in the city takes longer than 40 minutes by car, according to Google. There is no information available in Google Maps or other online source about

[1] Public schools that offer daycare, kindergarten and elementary education are run by the municipal government, whereas middle and high school years are offered in schools run by the state government; only a few private schools in our sample teach students from preschool to middle school. Indeed, 655 out of 1044 students did not attend the same school in the years of 2008 and 2012.

[2] We tested using the school attended in 2012 (instead of 2008) to anchor the location of the household. As expected, the corresponding instrument did not present a meaningful correlation with daycare/kindergarten attendance

[3] For students with missing values that studied in the same school in 2008 and 2012, it was considered the reported distance from school in 2012. Moreover, for students that only reported the mean of transportation but not the amount of time taken, we used the average time spend by children that went to the same school using the same mean of transportation as a *proxy*.

[4] Google allows to collect distance and duration of recommended routes between two adresses at a time, as explained in https://developers.google.com/maps/documentation/distance-matrix/intro

[5] Since we rule out the possibility of children that live in Cruz das Posses district go to preschool in the main part of Sertãozinho and vice-versa, we only considered distances between origin and destination within Sertãozinho and within Cruz das Posses, but not across them

routes using public transportation, nor school transport service. Some parents in the sample did not report the mean of transportation used.

Table 38: Number of routes collected from Google Maps Distance Matrix API, by mean of transportation

Duration (minutes)	Source: Google Maps			Not available on Google Maps		
	On foot	Bike	Car	Bus	School Transport	Not informed
0 - 10	56	412	656	-	-	-
11 - 20	145	373	189	-	-	-
21 - 40	361	48	61	-	-	-
41 - 90	252	70	-	-	-	-
91 - 210	90	-	-	-	-	-
TOTAL	904	903	906	-	-	-

The average speed by mean of transportation was obtained dividing the length of the route between two addresses by the estimated duration of the route, both provided by Google Maps. There was no substantial difference in the average speeds among routes departing from each the elementary school (which is reasonable, since Sertãozinho is a plain city with no region particularly difficult to reach), such that we consider the same average speed for all schools.

Table 39 displays the average speed applied over the duration (in minutes) from school reported by parents. For students that went to school on foot or by bike, we consider 80.1 and 284.4 meters per minute, respectively, which is the average for routes of all durations [6]. Routes by car provided by Google Maps presented a non-linear behavior, which is consistent with the idea that driving longer distances allow for higher speed, such that we consider different speeds according to the duration.

We made assumptions about the means of transportation whose routes are not provided by Google. Being Sertãozinho a small city, it was assumed that the time spent going by public bus is the same of going on foot for distances up to 40 minutes; routes that take 50 to 90 minutes would last 75% by bus, and for distances longer than 90 minutes walking, it takes 50% of the time by public transportation. It was also assumed that the duration of routes using school transport service is 75% than the duration of going by public transportation. To the 42 students that did not report the transport used, we used the same speed of students going on foot. In the last line we have the standard deviation of the calculated speeds from Google API raw data.

[6] There were tiny differences in the average speed according to the duration of walking routes, but it seemed more like measurement error, we opted to use the average calculated for all routes on foot. The average speed bicycling was increasing in duration of the route: an average 280 meters/minute (17 km/h) for routes that lasted up to 30 minutes and meters/minute (20 km/h) for routes of 31-60 minutes. However, we opted to consider the average of 282 meters/minute for all distances, since children about 8 years old would probably be tired after 30 minutes, at most maintaining the average speed.

Table 39: Average Speed (meters per minute) by mean of transportation and standard deviation

Duration (minutes)	Source: Google Maps			*Proxy* considered		
	On foot	Bike	Car	Bus	School Transport	Not informed
0 - 10	80.1	284.4	393.4	80.1	106.8	80.1
11 - 20	80.1	284.4	478.0	80.1	106.8	80.1
21 - 40	80.1	284.4	745.0	80.1	106.8	80.1
41 - 90	80.1	284.4	745.0	106.8	142.3	80.1
91 - 210	80.1	284.4	745.0	160.1	213.5	80.1
St. Dev. (all obs.)	4.3	41.8	186.3	-	-	-

With above information, we could calculate the greatest distance each student would be from school [7], and use it as the radius of the area each student would live in, according to provided information. This is how we build some of the shaded areas in gray in figure 4.

Table 40 summarizes the distribution of the calculated distance (in meters) that the children in our sample would be from school, using data and assumptions discussed above. We have that 17 children would live at 0 minutes from school, such that the household location would coincide with the school address; almost a third of our sample lived at a maximum distance of about 1 km from elementary school.

Figure 20 illustrates the role played by transportation mode. For instance, a child that takes 10 minutes walking would be at most 800 meters far from school, while the child that takes the same time by car could live at most 3,9 km away.

We limited the radius of the buffer to 8.5 km, which is enough to span all the urban area from any point in the city. We also consider the child lives inside of the urban limit, that nearly coincides with the street map.

E.2 Calculating the distance to nearest Early Childhood Center

Using QGis geographic information system we plotted the location of the elementary schools in 2008 and the corresponding areas where each students would live by, as exposed in the previous section.

Our instrument is the smallest distance from the border of the circle (which is the largest area the student would live in) to the nearest school offering daycare/kindergarten. We fall into four cases, being all of the corresponding distances also obtained from QGis:

[7] The speed considered over the length of the route, not over the straight distance of the two addresses. It is conservative in the sense that it is the greatest distance in a straight line, so it is the furthest each student could live from school.

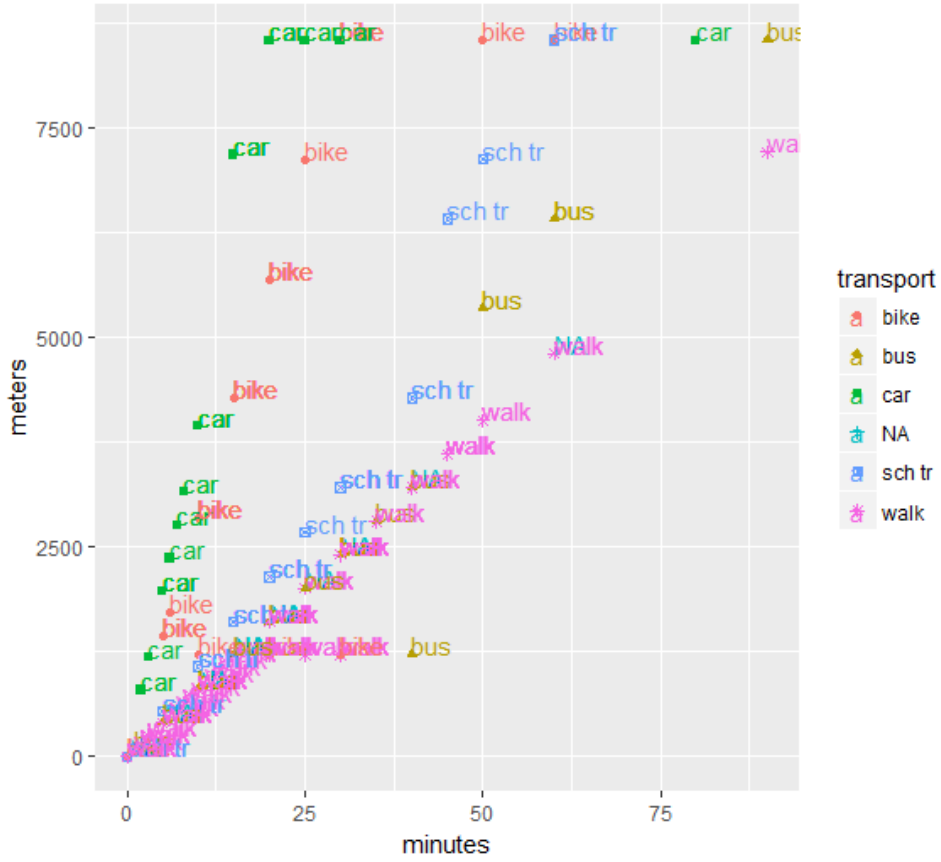


Figure 20: Estimated distance in meters from elementary school (vertical axis), reported distance in minutes (horizontal axis), by mean of transportation

Table 40: Distribution of calculated distance from household to elementary school attended in 2008

Distance from elementary school attended in 2008

Maximum distance (m)	Cumulative number of children	%
0	17	1.6%
250	28	2.7%
500	110	10.5%
750	128	12.3%
1000	266	25.5%
1500	470	45.0%
2000	615	58.9%
3000	757	72.5%
4000	882	84.5%
> 4000	1040	99.6%
Missing	4	0.4%
Total	**1044**	**100.0%**

1. *The area is a point*: this is the case if parents reported a distance of 0 minutes to school - the instrument is the distance to the nearest ECC (this is the case for 17 observations in the sample);

2. *There is no ECC in the area that the family lives*: we consider the distance to the nearest ECC out of the border (46 observations);

3. *There is a single ECC inside of the area*: it was considered the distance from the border of the area to the unique ECC (76 observations);

4. *Multiple ECCs inside of the area*: we compute the distance from the border of the area to the closest ECC inside of it (901 observations);

We also assumed that children that went to elementary school in Cruz das Posses lived in the district and could only have attend preschool in Cruz das Posses. We ruled out the possibility of the family travelling more than 12 km to take the child in daycare or kindergarten in Sertãozinho main city.

F – Household model of investment decision

Parents choose investment in the child in order to maximize the household utility function, taking in to account parents preferences, the perceived human capital of the child, the household characteristics and resources. For now, consider that each family has only the child, but the model can be extended to incorporate effects of siblings.

We consider that the household utility function is linear and separable, which is standard in this literature (CARNEIRO; HANSEN; HECKMAN, 2003; VERAMENDI; URZÚA, 2011). In each period t, the household derives utility from consumption (c_t) and from child's human capital in the following period, h_{t+1}, being ψ parents' altruism with respect to their child.

Parents do not observe the skill production function, but know that current investment (I_t) has a positive effect on child's human capital in $t+1$, i.e., $\frac{\partial h_{t+1}}{\partial I_t} > 0$. The cost of investing in the child formation in terms of utility depends on the household characteristics Z_t. We assume the costs to have a linear structure, i.e., $k(Z_t) = kZ_t$.

$$U(c_t, h_{t+1}, Z_t) = \beta\ ln(c_t) + \psi\ h_{t+1} - k(Z_t) \tag{F.1}$$

The household has an income stream w to be allocated between consumption and investment in the child, being p and q the prices of consumption and investment, that we assume to be constant overtime for simplicity.

$$p\ c_t + q\ I = w_t \tag{F.2}$$

In each period t the family problem is to choose c_t and I_t to maximize their utility subject to the above budget constraint. Since $h_{t+1} = g(\theta^C_{t+1}, \theta^N_{t+1})$, we compute the following first order conditions:

$$(FOC\ c_t) : \frac{\beta}{c_t} = \lambda p \tag{F.3}$$

$$(FOC\ I_t) : \psi\left[g'_{\theta^C_{t+1}} \frac{\partial \theta^C_{t+1}}{\partial I_t} + g'_{\theta^N_{t+1}} \frac{\partial \theta^N_{t+1}}{\partial I_t}\right] = \lambda q \tag{F.4}$$

$$(FOC\ \lambda) : p\ c_t + q\ I = w_t \tag{F.5}$$

From equations F.4 and F.5 we have that the optimal investment in the child depends on the relative price of investment and consumption, and of the magnitude of the partial derivatives with respect to investment.

In sum, we assume that parents have a desire to affect child's future human capital, and know that investing in the early childhood could affect the child's skills. The family has preferences and faces budget restrictions. The cost of investment can be both in terms of utility (letting a stranger take care of the child), monetary, if childcare is paid, or even in terms of child health - if the child attends a bad quality daycare, she can be ill more often (the cost would be capturing ECC quality, for instance).

We do not derive a closed form solution and a complete structural model because we to leave a flexible form of the production function (not restricting to be a linear or Cobb-Douglas). To draw further conclusions, we would also have to assume that the parents know the true skill production function.