Estimating a Dynamic Latent Factor Production Function for Skills and Health: Theory and Application to Children in Rural China

Jack Blundell

Balliol College, University of Oxford

May 2016

Submitted in partial fulfilment of the requirements for the degree of

Master of Philosophy in Economics

Word count = 27,405 (number of words on page 3 (315) x number of pages (87))

Abstract

A growing literature has demonstrated that skills, health and related investments in early years hold substantial predictive power for important outcomes well into adult life. An understanding of the mechanisms driving these relationships is not only of academic interest but vital for the effective design of social policy. In this paper I build on a rapidly growing field that seeks to explain child outcomes by treating skills, health and parental investments as unobserved factors for which we can hope to obtain related observable measures. These key factors are then combined with parental characteristics in the framework of a dynamic production function. The form and parameters of this function are the objects of interest. I follow a 'three step estimator', which is intuitively attractive and relatively computationally simple. My contributions are two-fold. First, I use new data from a randomised controlled trial in rural China to present additional empirical evidence for a Cobb-Douglas production function, in line with the existing literature. I use the estimated model to demonstrate the effectiveness of investments at different periods and explore implications for optimal policy. Second, using a simulation approach, I show that the distributional approximation that is commonly invoked may lead to a bias in parameter estimates. I then present a modification that gives a reduction in bias.

Acknowledgements

I thank my supervisors, Orazio Attanasio (University College London) and Steve Bond (Nuffield College, University of Oxford) for their invaluable advice and support throughout. I also thank Emily Nix and Sarah Cattan for access to their code, as well as Costas Meghir and Abhijeet Singh for their comments. Finally I thank Nele Warrinnier and Sean Sylvia for guiding me through the data, which was kindly provided by Rural Education Action Program, Stanford University (http://reap.fsi.stanford.edu/).

Contents

1 Introduction

A consensus is emerging that differences in cognitive and non-cognitive ability, health and numerous other important factors emerge early on in life and exhibit strong persistence into adulthood. In an era of high and growing inequality within many developed countries, the extent to which adult outcomes can be predicted by early indicators is not only of ethical concern but also represents a potentially substantial social cost. In less developed countries attempting to transition to advanced economies, there is concern over groups within society being 'left behind', unable to enjoy the growth on which others are able to capitalise due to the circumstances of their birth. As stated on the website of the Rural Education Action Programme (REAP): "It is possible that failing to educate and train poor rural children will jeopardize China's growth and transformation into a modern, knowledge-based economy."

Pioneering work by James J. Heckman and others, as discussed in depth in Carneiro and Heckman (2003), has shown that private and social returns to investment are exceptionally high in early years, a result which has significant implications for optimal social policy. This is particularly striking when compared to the low returns of various programmes later in life. Less is known about the channels through which this investment delivers these returns.

Early childhood has only been of interest to economists relatively recently, so it is important to look to other disciplines for insight. Drawing on developmental psychology and biology, the concept of 'critical periods', stages when the nervous system is particularly sensitive to external factors, is becoming increasingly recognised as important when looking at early years investments. As in the field of epigenetics,[1] discourse in Educational Economics has moved away from the traditional nature vs nurture debate to a more nuanced dynamic view, allowing for complex complementarities over time. While there is certainly a strong role for hereditary traits,

[1]Rutter (2012) presents a review of recent developments in gene-environment interactions.

previous work has suffered by confounding inherited ability with variation in childhood environment. To capture this complexity we must allow a certain degree of flexibility in functional form. This also means that the ideal data with which we estimate our models must have a rich time dimension.

A key consideration before embarking on any work of this type is to understand the benefits of taking a more 'structural' approach. Given that this often involves more time and computational power than a more 'reduced-form' approach (such as a value-added model), we must be clear of the advantage of such an approach. In defending my choice of the approach in this paper I would emphasise two clear benefits. Firstly, we are able to explore more flexible patterns of substitutability and complementarity than we would be able to with a typical linear-in-parameters reduced-form model. These may or may not be important, but taking a structural approach allows us to fully explore whether this is the case. Secondly, given that we are searching for 'deep' parameters that are policy-invariant, we are able to conduct policy experiments. Although the use of experiments is growing, the set of potential policies will always dwarf what can be learned through experimental design.[2]

The evolution of a 'standard' way of tackling the problem of estimating a dynamic latent factor production function in its entirety is still very much in its nascent stage, hence the need for articles such as this one.

This thesis is structured as follows. Firstly, I explain its relation to various strands of existing literature, both historical and recent. Secondly, I give an overview of the theory behind the method I am investigating throughout. In Section 4 I estimate the model using new data from REAP, an intervention in China. I present results that are broadly in line with other estimates, which given the contrast in settings is intriguing in itself. Using my estimated production functions, I am able to briefly explore the implications for optimal investment. Next, I present simulations demonstrating that the distributional assumption currently used in part of

[2]Keane (2010), Wolpin (2013) and Rust (2010) present a full overview of the arguments in favour of structural modelling. Todd and Wolpin (2003) give arguments specific to the estimation of Education Production Functions.

the literature may result in a bias that can be substantial. I also show how the assumption can be modified to reduce the bias. In the final section I summarise my contributions, discuss some limitations of my analysis and indicate the many directions in which I believe further research is needed.

2 Relation to the literature

The existing literature is best split into several groups: early work on life cycle skill formation,[3] research on the multi-dimensionality of human capital, dynamic complementarities, evidence on interventions and the recent work on dynamic latent factor production functions.

2.1 Early work on life cycle skill formation

Historically, labour economists have viewed human capital as having two distinct components, innate ability and acquired skills.[4] The former is determined by inherited traits and a significant random component, and in the standard model is treated as unobservable. This is encapsulated in the Ben-Porath (1967) model, which describes a set of 'initial endowments' that a child receives at birth. The latter, which until more recently has been the focus of work on human capital, is modelled as a choice variable and determined by numerous child, parental and institutional characteristics (Becker and Tomes, 1979).

Often, cumulative years of schooling has been adopted as a measure of skill, as most famously seen in the Mincer (1974) earnings equation. As shown in Lemieux (2003), the Mincer equation to an extent continues to offer a good fit empirically. However, as explained in Bowles et al (2001), there remains ample unexplained variation in wages. Over half of the variance in the natural logarithm of wages

[3] In their Handbook of the Economics of Education chapter, Cunha et al. (2006) present a wide and detailed review of the literature on life cycle skill formation.

[4] The term 'human capital' is most famously coined by Becker (1964), but increasingly became a topic of academic discussion in the 1940s. Becker's contribution and historical context is described in Teixeira (2014).

cannot by predicted by conventional demographic and skill variables. This motivates further investigation into whether we can more accurately measure acquired skills, and furthermore into what we miss by subscribing to the traditional dichotomy between ability and skills.

One broad subset of literature looks more closely at the inputs that feed into a child's education. An early foray towards an input-based approach was the Coleman Report (Coleman et al., 1966). As outlined in Hanushek (2003), this paved the way to a consideration that inputs can be combined in a production function to form outcomes, much as a firm takes inputs and uses available technology to produce outputs. This bears similarities in approach to the more recent production function literature. Much of this literature looked at schooling inputs as opposed to inputs from parents made at home. Proposed inputs were teacher quality, school resources and class sizes, which in the more recent literature would be considered as measures for a latent investment variable. Additionally, often there was little concern for whether estimates were truly causal, which is clearly limiting if studies are taken as guides for policy. Much of the debate centred around specific policy issues such as whether a reduction in class sizes has an effect on student achievement.

2.2 The multi-dimensionality of human capital

While ability has traditionally been viewed both as a scalar and as fixed over time, numerous studies have now shown there to be a variety of distinct factors that interact to generate outcomes. This has motivated the interpretation of ability as a multi-dimensional object.

Rubinstein and Heckman (2001) investigate the response of employers to the General Educational Development (GED) test. They argue that the GED is a mixed signal, and in fact demonstrates a lack of 'non-cognitive' ability, which is penalized by the labour market. Further evidence for the importance of non-cognitive skills, which might include traits such as self-discipline, motivation and socio-emotional skills, on labour market and behavioural outcomes is presented in Heckman, Stixrud,

and Urzua (2006). Taking a latent factor approach, they find that cognitive skills are important for outcomes, as is the standard view, but that non-cognitive skills impact on both acquisition of schooling and on labour market performance for a given level of schooling. Roberts et al. (2007) provide further evidence, demonstrating that for many outcomes, including mortality and occupational attainment, personality traits rival cognitive ability for predictive power.

Using Australian data, Fiorini and Keane (2014) explore how the allocations of a child's time affects skill development. They find that spending time pursuing educational activities leads to higher cognitive skills, and that this is particularly true if these activities involve parents. Non-cognitive skills on the other hand appear to be influenced primarily by parenting style.

Since Grossman's (1972) model of health production, economists have started viewing health as a type of human capital, much as we view cognitive and non-cognitive skills. Analogous to the literature on education, many recent studies have demonstrated the importance of health in early years on a variety of later outcomes. As reviewed in Behrman (1996), early health and cognitive skills appear to interact to generate outcomes. A recent study demonstrating this using US birth and schooling data is Figlio et al. (2013). Behrman and Rosenzweig (2004) discuss the 'returns to birthweight', suggesting that interventions affecting birthweight could have a substantial effect on global inequality. Black, Devereux, and Salvanes (2007) provide further evidence of the importance of birthweight on IQ, earnings and education using Norwegian data. In Canada, Oreopoulos et al. (2006) show that infant health is a strong predictor of labour market and educational outcomes. Lucas et al. (1998) find that a nutritional intervention in neonatal units in the UK had a significant effect on IQ test scores at age 7-8. Bharadwaj et al. (2013) use data from Norway and Chile to show that additional medical care at birth leads to higher test scores at school.

Much of the more recent work in health has focused on pre-natal conditions. For example, Almond, Edlund, and Palme (2009) use the natural experiment of the

Chernobyl nuclear disaster to show that pre-natal exposure to radiation in Sweden can have a long-term effect on academic performance, and that this is highly concentrated among children of parents with low levels of schooling. Black, Bütikofer, Devereux, and Salvanes (2013) present further evidence of the impact of low-dose nuclear radiation on children in utero, using variation in Norway resulting from nuclear weapon testing. Almond (2006) uses the 1918 flu epidemic in the US to show that in utero conditions matter for physical disability as an adult, as well as income, socioeconomic status and welfare reliance. The dynamic production function approach is equipped to include pre-natal conditions. It is plausible that if the right data is available, the in-utero period can be included as an initial period in the dynamic model. Indeed, to capture the complete picture of child development, this seems necessary.

2.3 Dynamic complementarities

In addition to the multi-dimensional nature of skills, in recent years greater attention has been paid to the role of dynamic complementarities. Becker and Tomes (1986) recognise the importance of early investments, but assume that early investments are proportional to later investments. Additionally, by treating early investments as occurring within just a single period, they implicitly assume that investments in this period are all perfect substitutes with one-another.

The assumption of perfect substitutability of investments across time is out of line with evidence from the developmental psychology literature. As is well known with language acquisition (Newport, 2003), human learning is structured around so-called sensitive and critical periods. These are stages in development when certain neural circuits are most plastic and hence most easily influenced by outside forces. As discussed in Knudsen et al. (2006), failure to understand and exploit these periods results in a large number of adverse outcomes for the individual in question. At the societal level, the social efficiency cost associated with misguided child investment profiles is substantial (Heckman, 2008).

2.4 Evidence on interventions

There is a burgeoning amount of evidence on childhood interventions targeting cognitive and non-cognitive skill development, the main results of which I present here.[5] I focus on studies which have had a long-run follow-up period, as this has been shown to be important in capturing the nature of any treatment effects. Ignoring the long-run effects of programs often understates the impact of early interventions.

The Perry Preschool Program is perhaps the best-known such intervention. The program provided preschool 5 days a week and home visits to black children in Michigan who performed poorly on an initial IQ test. The two-year program had large initial effects on IQ, but these faded out so that by the age of 10, the treatment and control groups had the same average IQ. This fadeout has been witnessed in numerous interventions (Bailey et al., 2015), and has been argued as showing that early interventions are ineffective.[6] However, as shown in Heckman et al. (2010), the long-run effects on earnings, criminal activity and welfare dependance were significant. Heckman, Pinto, and Savelyev (2013) investigate the channels through which these later benefits are realised, and find that it appears to be primarily through non-cognitive skills.

Another well-known US study is Project STAR. This saw the randomisation of Tennessee students and teachers into kindergarten classrooms through to Third grade. Chetty et al. (2011) look at how variation in class size and teacher experience affects later schooling outcomes. They and others find a positive effect on test scores in the short term, but full fadeout after several years. However, by linking students to administrative records, it is shown that there were in fact long-lasting impacts on earnings later in life. As in the Perry Preschool Program, these differences can be attributed to non-cognitive skills.

As well as those in the US, many interventions have taken place in less developed countries. The Jamaica Supplementation study consisted of a randomised

[5] Almond and Currie (2011) present a detailed review of evidence from many early-years interventions.

[6] Jensen's 1969 article in the Harvard Educational Review caused controversy by making this argument, based on the efficacy of the 'Head Start' Programme.

nutritional and stimulation program for stunted children. As shown in Walker et al. (2005), the stimulation program had large and long-lasting effects, and the nutritional program had smaller effects. The benefits came both through cognitive and non-cognitive channels.

These results have implications for the classic equity-efficiency tradeoff that appears in many fields of economics (Okun, 1975). As discussed in Heckman and Masterov (2007), the implication of empirical work on early interventions is that investing in young disadvantaged children both reduces inequality and improves efficiency. This is a case when motivations of fairness and social justice coincide with utilitarian economic goals.

The above arguments suggest first that skills are multi-dimensional and second that dynamics are important. The next relevant question to ask is then whether the different dimensions of skill are associated with different sensitive periods. As reviewed in Kautz et al. (2014), cognitive skills seem to become fixed reasonably early.[7] However, non-cognitive skills are thought to be malleable well into adolescence. Although there is less evidence than for early interventions, Kautz et al. (2014) describe numerous workplace-related schemes that improve outcomes via non-cognitive skills such motivation, optimism and autonomy. As reviewed in Heckman et al. (1999), job training programs in adulthood tend to have small private and social gains, consistent with falling returns to investments over the lifecycle.[8]

The above evidence suggests that 'skill begets skill', and points towards two clear mechanisms that skill attainment appears to follow. Firstly, skill attainment in one period raises that in the next. This is described as 'self-productivity'. Secondly, investments in different periods are complementary. This means that investments that are un-matched in other periods are less productive. The recent dynamic latent

[7]This is disputed in recent work by Carlsson et al. (2015), who find cognitive skills are affected by schooling and hence are still malleable up to age 18. The authors suggest that the type of tests typically used as cognitive indicators lends itself to finding bigger results on malleability at early ages, and that it is disputable whether these truly represent cognitive ability.

[8]This is vividly captured in the 'Heckman Curve' (http://heckmanequation.org/content/resource/heckman-curve).

factor production function literature seeks to create a tractable, structural model that is consistent with these two features.

2.5 Dynamic latent factor production functions

A recent breakthrough article in dynamic latent factor production functions is Cunha and Heckman (2007). Here a simple framework is set out that encapsulates many of the empirical findings discussed in the previous sections. As will be described in detail in Section 3, the model comprises of a Constant Elasticity of Substitution production function with current stocks of skills depending on past stocks of skills, investments and parental characteristics.

The model is tractable and attractive in its potential to explain empirical findings, however it is not clear how best to go about estimation. Building on the work of Todd and Wolpin (2003), Cunha and Heckman (2008) exploit covariance restrictions to achieve identification of the production function. As is discussed in the paper, it is difficult to distinguish empirically between investments in different types of skills, hence investment within each period is treated as one latent input factor. The key weakness of this paper is that a linear structure is assumed, so that there is perfect substitutability between early and late investments. This is out of line with the stylised facts discussed previously.

While non-linear models are clearly needed, applying such models to variables that are observed with measurement error results in an error term that is not additively separable. Hence standard Instrumental Variables techniques cannot be used. Schennach (2004) presents a solution to this problem using repeated observations and a useful property of the Fourier transform. Cunha et al. (2010) draw on this to achieve identification in a more general nonlinear technology than that of Cunha and Heckman (2008). To estimate the model, an Unscented Kalman Filter is applied, an algorithm that uses a series of measurements over time to produce estimates of unknown variables. The data they use is on 0-14 year old children of the NSLY/79 sample. Much attention is then dedicated to the elasticity of sub-

stitution parameter, which shows the degree of complementarity between different inputs. The authors find that self-productivity becomes stronger as children become older, as does complementarity between cognitive skills and investment. This is not true of non-cognitive skills. Hence this evidence is consistent with empirical findings that interventions on disadvantaged children operate most successfully though non-cognitive channels.

Two recent papers have proposed an alternative estimation method to address the same problem as in Cunha et al. (2010). This 'three step estimator'[9] is the focus of my paper. As the method itself is described in detail in Section 3 of this paper, here I will discuss the empirical results only.

Attanasio, Meghir, and Nix (2015b) estimate production functions for cognition and health using data from the Young Lives Survey for India, over a similar age range to that of Cunha et al. (2010). They find that investments are important and that inputs are complementary. Notably, health is important in determining later cognition. This is in line with evidence such as Figlio et al. (2013) in the US context, and with the long-term effects of a nutrition intervention in Guatamala shown in Hoddinott et al. (2008).

Attanasio, Cattan, Fitzsimons, Meghir, and Rubio-Codina (2015a) adopt the same method in a novel setting. They use the production function framework to evaluate the channels through which a randomised early childhood intervention in Colombia led to gains in cognitive and non-cognitive ('socio-emotional') skills. In a short-run follow-up study, Attanasio et al. (2014) demonstrate that the psychosocial stimulation element of the intervention had substantial effects on language and cognitive development. They find that gains can be fully attributed to increased investments, and once again a strong role for complementarity.

Another innovation in Attanasio et al. (2015a) is the separation of investments into time and commodities. The authors show that material investments are more associated with cognitive gains, and time investments with socio-emotional skills.

[9]The properties of the three step estimator are being formalised in Cattan, Nix, and Stouli (2016).

Alongside the empirical section of my paper, this is also one of the few examples in the production function literature that looks at children in the first three years of life. While the Colombian data is richer in the cross-sectional sense than the data for China used here, it is limited to only the two periods (treatment and follow-up), whereas I am able to exploit a richer time dimension to give a series of production functions.

In all the papers described in this section, the authors are unable to reject a Cobb-Douglas technology of skill formation, corresponding to an elasticity of substitution of one. Given the variety of different settings in which these functions have been estimated this is a striking result, and one that in part motivates the simulation evidence presented in Section 5.

3 Methodological framework

In this section I outline the approach that will be explored in Sections 4 and 5. Of the two approaches to estimation in the existing literature, I follow that of Attanasio et al. (2015b) as opposed to Cunha et al. (2010). Much of the conceptual framework underlying these approaches was developed by Cunha and Heckman (2007). Since the method is new, I present it here in some detail, and refer the reader to Attanasio et al. (2015b) for a full exposition. Readers familiar with the method may prefer to move ahead to Section 4.

3.1 Model of skill formation

Each child's stock of human capital, which can consist of cognitive skills, non-cognitive skills and health, is considered over a discrete set of periods $t = 0, 1, 2..., T$.[10] As in Attanasio et al. (2015b) I abstract from non-cognitive skills and focus purely on cognitive skills and health, which I denote by $\theta_{c,t}$ and $\theta_{h,t}$ respectively at each pe-

[10] In practice, the periods themselves are typically chosen on the basis of data availability. It is an open question as to how best to choose these periods, and how choice of period length affects parameters.

riod t. The exclusion of non-cognitive skills is not due to relative unimportance, but is primarily driven by data requirements and the additional complexity of modelling all three components simultaneously.

At birth a child is endowed with a stock of human capital. Although in many instances it would be possible to explore, here I do not investigate the determinants of this initial stock, only the development of further human capital conditional on these starting values. In later periods, the stock of each of these 'outcome' variables in each period $t+1$ is determined by stocks in the previous period t and three further input factors. Stocks at t then enter the model both as outputs at period t and as inputs at period $t+1$.

The first of the additional input factors is investment in the child in the previous period denoted by I_t. This investment may be divided into different types of investment, and represents contributions from both parents and government. We might think of investment here as anything from diet to time spent interacting with the child.

The two remaining inputs are parental health θ_{ph} and parental cognitive skills θ_{pc}, which are considered fixed over time. Given that investment is accounted for, these factors primarily capture heritability. Note that allowing parental health and cognition to enter the production function directly in all periods allows us to test whether these characteristics only affect a child's initial starting conditions, or whether inherited characteristics have additional impacts in later periods. Attanasio et al. (2015b) for example find that parental cognition is significant only the earliest production function, whereas parental health continues to have an impact in later periods. One explanation for this is that health is influenced by many lifestyle factors, and the habits of parents are likely to influence those of their children over an extended period. This explanation works for children of the age range in their article (5-15 years), but is less plausible in the empirical setting in Section 4 of this paper (0-2 years).

Finally, I allow for unobserved shocks in the production processes for cognitive

skills and health, denoted respectively by $u_{c,t}$ and $u_{h,t}$. In our notation, shocks indexed by t affect the output factor at time $t+1$. This is against the convention in time-series econometrics but is in-keeping with the previous production function literature.

I can formalise this as:

$$\theta_{c,t+1} = f_t^c(\theta_{c,t}, \theta_{h,t}, I_t, \theta_{ph}, \theta_{pc}, u_{c,t}) \tag{1}$$

$$\theta_{h,t+1} = f_t^h(\theta_{c,t}, \theta_{h,t}, I_t, \theta_{ph}, \theta_{pc}, u_{h,t}) \tag{2}$$

The fundamental question of interest is the form taken by equations (1) and (2), or how inputs combine to generate outputs. Note that both functions are indexed by t, so are allowed to vary over time. This is important if we are to get a full picture of development throughout childhood, and imposing constant parameters over time would limit the range of questions we could hope to answer. As discussed in Cunha and Heckman (2007), varying the form of these functions has different implications for the technology of skill acquisition, which can include features such as dynamic complementarity and self-productivity.

The dominant functional form assumption in the current literature is that of a Constant Elasticity of Substitution (CES) function with constant returns to scale. Adapting equations (1) and (2) to this setting gives the following:

$$\theta_{c,t+1} = (\delta_{c,t}\theta_{c,t}^{\rho_t} + \delta_{h,t}\theta_{h,t}^{\rho_t} + \delta_{I,t}I_t^{\rho_t} + \delta_{ph,t}\theta_{ph}^{\rho_t} + \delta_{pc,t}\theta_{pc}^{\rho_t})^{1/\rho_t} exp(a_{c,t} + u_{c,t}) \tag{3}$$

$$\theta_{h,t+1} = (\alpha_{c,t}\theta_{c,t}^{\xi_t} + \alpha_{h,t}\theta_{h,t}^{\xi_t} + \alpha_{I,t}I_t^{\xi_t} + \alpha_{ph,t}\theta_{ph}^{\xi_t} + \alpha_{pc,t}\theta_{pc}^{\xi_t})^{1/\xi_t} exp(a_{h,t} + u_{h,t}) \tag{4}$$

$$\delta_{c,t} + \delta_{h,t} + \delta_{I,t} + \delta_{ph,t} + \delta_{pc,t} = 1$$

$$\alpha_{c,t} + \alpha_{h,t} + \alpha_{I,t} + \alpha_{ph,t} + \alpha_{pc,t} = 1$$

Shocks $u_{c,t}$ and $u_{h,t}$ are included in such a way that when logs are taken, they enter additively. I have also added total factor productivity (TFP) terms $a_{c,t}$ and

$a_{h,t}$, which are constant across individuals and can be estimated along with parameters. In the Cobb-Douglas case, these correspond to the intercept of a log-linear specification, as will be shown shortly. These ensure that the shock terms have an expected value of zero.

The elasticity of substitution for each equation is given by:

$$\eta_{c,t} = \frac{1}{1-\rho_t} \tag{5}$$

$$\eta_{h,t} = \frac{1}{1-\xi_t} \tag{6}$$

This elasticity governs substitution behaviour between different inputs and is therefore of key importance.

This is a tractable and convenient functional form assumption for a number of well-known reasons. Chiefly, it nests within it a variety of cases of particular interest. Ignoring the shock term, when $(\xi_t, \rho_t) = 1$ the function implies perfect substitutability of inputs. This would mean that we can perfectly compensate for example a lack of cognitive ability in one period with high investment in that period. Alternatively, as $(\xi_t, \rho_t) \to -\infty$, we have a Leontief function, corresponding to perfect complementarity. This is the case where no amount of investment could compensate for low cognitive ability in a period. These extremes of perfect substitutability and perfect complementarity are out of line with empirical evidence and basic intuition on skill development.

In the intermediate case in which $(\xi_t, \rho_t) \to 0$, the function reduces to a Cobb-Douglas, which can be written in the following log-linear form:

$$\begin{aligned} ln(\theta_{c,t+1}) = \delta_{c,t} ln(\theta_{c,t}) + \delta_{h,t} ln(\theta_{h,t}) + \delta_{I,t} ln(I_t) \\ + \delta_{ph,t} ln(\theta_{ph}) + \delta_{pc,t} ln(\theta_{pc}) + a_{c,t} + u_{c,t} \end{aligned} \tag{7}$$

$$ln(\theta_{h,t+1}) = \alpha_{c,t}ln(\theta_{c,t}) + \alpha_{h,t}ln(\theta_{h,t}) + \alpha_{I,t}ln(I_t) + \alpha_{ph,t}ln(\theta_{ph}) + \alpha_{pc,t}ln(\theta_{pc}) + a_{h,t} + u_{h,t} \quad (8)$$

This functional form implies diminishing marginal returns to each input, and represents a compromise between the aforementioned notions of perfect substitutability and perfect complementarity.

Although there are useful qualities of the CES assumption, it implies costly limitations on the interactions between inputs. In particular, it leads to an elasticity of substitution between all pairs of inputs that is identical. It is difficult to argue that the elasticity of substitution between parental health and a child's own past health should be the same as that between parental health and a child's own past cognitive ability. The translog (transcendental logarithmic) production function, a generalisation of the Cobb-Douglas in which squares and cross-products of the log-inputs are included, does not suffer from the same criticism.[11] This is yet to be explored in the current literature.

If we were able to treat $u_{c,t}$ and $u_{h,t}$ as independent of all inputs and could observe all the relevant variables, equations (3) and (4) could be estimated simply using Non-linear Least Squares, or even Ordinary Least Squares on equations (7) and (8) if we assumed the Cobb-Douglas form. However, neither of these conditions is likely to hold. Each of these complications and the proposed solution is discussed in the following sections. I pay particular attention to the issue of unobservability of factors, as this is the aspect that I explore with simulation evidence in Section 5.

[11] A different approach which would offer a more intuitively appealing way of including parental factors (particularly in later periods) would be to view parental factors as affecting the productivity of inputs. This would be possible by allowing the coefficients on each input factor to vary with each child's parental characteristics. This captures interaction effects such as investments having differential impacts depending on parental characteristics. We may imagine for example that a child with parents of higher cognitive ability may in general see greater cognitive gains from investments such as being read to.

3.2 Investment

Following the modelling strategy in Cunha and Heckman (2007) and Attanasio et al. (2015b), I assume that parents are altruistic towards their children. Parents decide how much to invest based on current and future returns, trading off their child's future wellbeing against their own consumption. Optimal investment will then depend on all other inputs in the production function, including the shocks $u_{c,t}$ and $u_{h,t}$. Although these shocks are treated as unobserved by the researcher, they may be at least partially observed by parents and hence influence investment decisions. The intuition behind this is that parents may choose to invest to compensate for a child's lack of ability, or alternatively they may invest more if their child is high ability as they expect higher returns to this investment. For health, we would imagine that investments will both affect the health of the child and will be influenced by periods of good or bad health. If investment is generalised to include government programmes, we might also expect this to depend on the stocks of child cognition and health, with support targeted at those most in need or those who suffer bad shocks. Investment is hence likely to be endogenous and non-linear least squares on (3) and (4) would lead to inconsistent estimates.

Investment will also depend on their own budget constraint, which importantly does not appear directly in the production function. An implication of the production function framework I adopt is that elements of the budget constraint should not affect skill formation directly, only via their effect on investment. Resources and prices then do not affect production directly once all other inputs are taken into account. Provided these elements are predictive of investment, which seems highly likely and is testable, we have then identified potential instruments.

Ideally, both price and resource information would be used, however since prices are not available in the data on which the empirical section of this paper is based, I focus here on household wealth. Letting Y_t represent wealth in period t, I follow the literature in assuming that we can approximate the determinants of investment

by a log-linear form:

$$ln(I_t) = \gamma_0 + \gamma_1 ln(Y_t) + \gamma_2 ln(\theta_{c,t}) + \gamma_3 ln(\theta_{h,t}) + \gamma_4 ln(\theta_{ph}) + \gamma_5 ln(\theta_{pc}) + v_{I,t} \quad (9)$$

Here the error term $v_{I,t}$ is allowed to be correlated with $u_{c,t}, u_{h,t}$, taking into account the fact that parental investments respond both to observables and the shocks which are unobserved to the econometrician. For wealth to be a relevant instrument, we require $\gamma_1 \neq 0$, which is a testable assumption.

For wealth to be a valid instrument we also require the exclusion restriction. This is essentially that wealth is not related to shocks to the production process. There are potential issues with this assumption. One could argue that unobserved shocks to child skill development could be correlated with available measures of wealth, for example by parents working more if a child becomes sick. This is a legitimate concern and suggests that better measures of wealth will be those that are not easily adjustable to shocks. Attanasio et al. (2015a) raise this concern and use tests of over-identifying restrictions to demonstrate the validity of wealth as an instrument in their setting.

Control Function methods rely on similar identification conditions to two stage least squares (2SLS) and coincide with 2SLS in a linear model. The approach typically requires less assumptions than would be required for maximum likelihood and is computationally simple.[12] Formally, we can give the key exclusion restriction assumptions for the control function strategy as:

$$E(u_{c,t}|\theta_{c,t}, \theta_{h,t}, I_t, \theta_{ph}, \theta_{pc}, Y_t) = \kappa_c v_{I,t} \quad (10)$$

$$E(u_{h,t}|\theta_{c,t}, \theta_{h,t}, I_t, \theta_{ph}, \theta_{pc}, Y_t) = \kappa_h v_{I,t} \quad (11)$$

The conditional expectation of the residuals from each production function are assumed to be linearly related to the residuals from the investment equation. Here,

[12] A recent overview of the theory and usage of Control Function methods is presented in Wooldridge (2015).

$v_{I,t}$ is the 'control function' and can be estimated by performing OLS on the investment equation (9). This estimate $\hat{v}_{I,t}$ is then included in the production functions as follows:

$$\theta_{c,t+1} = (\delta_{c,t}\theta_{c,t}^{\rho_t} + \delta_{h,t}\theta_{h,t}^{\rho_t} + \delta_{I,t}I_t^{\rho_t} + \delta_{ph,t}\theta_{ph}^{\rho_t} + \delta_{pc,t}\theta_{pc}^{\rho_t})^{1/\rho_t} exp(a_{c,t} + u_{c,t} + \hat{v}_{I,t}) \quad (12)$$

$$\theta_{h,t+1} = (\alpha_{c,t}\theta_{c,t}^{\xi_t} + \alpha_{h,t}\theta_{h,t}^{\xi_t} + \alpha_{I,t}I_t^{\xi_t} + \alpha_{ph,t}\theta_{ph}^{\xi_t} + \alpha_{pc,t}\theta_{pc}^{\xi_t})^{1/\xi_t} exp(a_{h,t} + u_{h,t} + \hat{v}_{I,t}) \quad (13)$$

Provided our assumptions on the determinants of investment are valid, non-linear least squares on (12) and (13) gives consistent estimates of all parameters as discussed in Wooldridge (2015).

3.3 Unobservability of factors

As well as the endogeneity of investment, a fundamental difficulty with estimating equations (3) and (4) is that the variables of interest cannot be directly observed. There is no single measure that can be argued as representing cognitive ability in its entirety, nor is there an analogous measure for health.[13] Investment is perhaps even more difficult to envision as being captured by a single measure.

The best we can hope to achieve is to combine a set of measures which contain information on the latent factor we are hoping to identify. This is a familiar problem in the Psychometrics literature, which presents many potential strategies to extract factors from measures. The approach adopted here is closer to that of a classical measurement error system than to conventional factor analysis.

First we define θ be the vector of all factors from all time periods:

$$\theta := ((\theta_{c,t})_{t=1}^T, (\theta_{h,t})_{t=1}^T, (I_t)_{t=1}^T, \theta_{pc}, \theta_{ph}, (Y_t)_{t=1}^T)$$

Letting M be a vector of observable measurements, we assume that measures are linear combinations of the log of latent factors and a matrix of normal measurement

[13] As discussed later in this paper, parental cognitive ability is often assumed to be observed without error, primarily due to data availability.

errors ϵ. We then assume that factors and measures are related by the factor loading matrix Λ, i.e.

$$M = A + \Lambda ln(\theta) + \epsilon \tag{14}$$

where

$$\epsilon \sim N(0, \Sigma^{\epsilon})$$

Log-factors are normalised so that they are mean zero, meaning that A is the vector of measurement means.[14] Letting $\tilde{M} := M - A$ be a vector of de-meaned measurements, we then arrive at the following:

$$\tilde{M} = \Lambda ln(\theta) + \epsilon \tag{15}$$

In our analysis, we assume that Σ^{ϵ} is diagonal, which corresponds to measurement errors that are independent across measures. This simplifies the discussion that follows, but can be relaxed quite simply.

Additionally, we assume a dedicated measurement system, which means that each measure is only (directly) influenced by one factor. Normalising the loading on one measure per factor to unity, a typical loading matrix Λ will then be of the following form:[15]

$$\Lambda = \begin{bmatrix} 1 & 0 \\ \lambda_{1,2} & 0 \\ \lambda_{1,3} & 0 \\ 0 & 1 \\ 0 & \lambda_{2,2} \\ 0 & \lambda_{2,3} \end{bmatrix}$$

[14] This normalisation influences the interpretation of the production function. As each log-factor is normalised to zero, factors capture for example the cognitive skills of a child relative to others of the same age, as opposed to relative to those of that same child in the baseline period. Another effect of the normalisation is that the TFP terms $a_{c,t}$ and $a_{h,t}$ become redundant.

[15] This specific example has two factors and three measures per factor and is purely illustrative.

Note that the setup here is general enough to include in the vector θ factors that are observed without error. In this case, the diagonal component of Σ^{ϵ} corresponding to the factor measured without error is constrained to be zero.

The key assumption made is that the measurement system follows a mixture of two normal distributions.[16] We can denote this by:

$$\tilde{M} \sim \tilde{\tau}\tilde{f}^{a} + (1-\tilde{\tau})\tilde{f}^{b} \tag{16}$$

$$\tilde{f}^{i} = N(\tilde{\mu}^{i}, \tilde{\Sigma}^{i}) \qquad i = a, b$$

Here, we can think of the vector of measures for each child as being drawn with probability $\tilde{\tau}$ from component a and with probability $1-\tilde{\tau}$ from component b, where each component is a multivariate normal distribution. Note that $\tilde{\tau}$ is a scalar and is not allowed to vary across individuals or measures. Since $\tilde{M}$ is mean zero, the weighted sum of $\tilde{\mu}^{a}$ and $\tilde{\mu}^{b}$ in which the weights are $\tilde{\tau}$ and $1-\tilde{\tau}$ will equal zero in expectation, although $\tilde{\mu}^{a}$ and $\tilde{\mu}^{b}$ themselves in general will not equal zero.

Under distributional assumption (16) and measurement-factors assumption (15) the log-factors also follow a mixture of two normals:

$$ln(\theta) \sim \tau f^{a} + (1-\tau) f^{b} \tag{17}$$

$$f^{i} = N(\mu^{i}, \Sigma^{i}) \qquad i = a, b \tag{18}$$

The beauty of assumption (16) is that we are able to link the mixing parameter, means and variances with the following linear equations:

$$\tilde{\tau} = \tau \tag{19}$$

[16] Mixture distributions are often used in cases when it is believed that the data-generating process can be split into two or more sub-processes. However, this is not the motivation in this setting. Instead, by choosing a mixture distribution we hope to allow for a high degree of flexibility while maintaining a distribution with convenient analytical features (see Geweke and Keane (1997) for a generalisation of the probit model based on similar motivations). As the number of components within the distribution increases, it is possible to approximate any distribution (Norets and Pelenis, 2012).

$$\tilde{\mu}^i = \Lambda \mu^i \qquad i = a, b \tag{20}$$

$$\tilde{\Sigma}^i = \Lambda \Sigma^i \Lambda' + \Sigma^\epsilon \qquad i = a, b \tag{21}$$

The given linking equations allow for any matrix of factor loadings Λ and any degree of correlation between the measurement errors (Σ^ϵ need not be diagonal). The linking equations lead naturally to a strategy by which we can estimate the distribution of factors from the distribution of measures. These also extend in the natural way to mixtures of more than two normals, which will be explored in later sections.

As outlined in Attanasio et al. (2015b), there are two assumptions required for identification of the factor loadings and the joint distribution of factors and measurement error (parameters $\Lambda, \Sigma^a, \Sigma^b, \Sigma^\epsilon, \mu^a, \mu^b, \tau$). Firstly, measurement errors must be independent of the latent factors. Secondly, if the number of unobserved factors is K we must have at least 2K + 1 measures, with at least two for each factor. This is sufficient for non-parametric identification.

3.4 Estimation

The previous subsections primarily set out the issues surrounding identification of the production function and the latent inputs and outputs. The estimation of equations (3) and (4) brings together these discussions into an intuitive multi-step process that is relatively straightforward to apply to data.

3.4.1 Estimating the distribution of measures

Statisticians have been utilising mixture distributions for at least a century (Pearson, 1894), but practical applications were limited until recently by the computational complexity involved in estimation. The most common way of estimating the parameters of a mixture distribution is the Expectation Maximisation (EM) algorithm, originally formalised in Dempster et al. (1977). EM is an iterative procedure that coincides with maximum likelihood, and is best suited to situations in which there is

some unobserved latent variable. In these situations, standard maximum likelihood is infeasible due to the missing element. In the case of a mixture distribution, the unobserved latent variable is which mixture component each observation is drawn from. Further details of how the EM algorithm works in this setting are given in Appendix A.

The EM algorithm provides (estimates of) the means, variances and mixing parameter from the distribution of measures:

$$(\tilde{\mu}^a, \tilde{\mu}^b, \tilde{\Sigma}^a, \tilde{\Sigma}^b, \tilde{\tau}) \tag{22}$$

where a and b are the two components the mixture distribution.

3.4.2 Minimum distance: from measures to factors

The next stage of estimation is to link the estimates of the parameters in (22) to the parameters of the latent factor distribution. This is achieved by applying the 'linking' equations (19), (20) and (21), minimising the squares of sample deviations from these population counterparts.

The criterion function for when we assume a mixture of two normals a and b is given in equation (23):

$$J := \sum_{k=a,b} \sum_{(i,j)|i\geq j} (\tilde{\Sigma}^k_{i,j} - (\Lambda\Sigma^k\Lambda')_{i,j} - \Sigma^\epsilon_{i,j})^2 + \sum_{k=a,b} \sum_i (\tilde{\mu}^k_i - (\Lambda\mu^k)_i)^2 \tag{23}$$

Here, $A_{i,j}$ corresponds to the entry in row i and column j of matrix A, and B_i corresponds to the i'th element of vector B. Summing over (i,j) corresponds to adding up all entries in the relevant matrix or vector, and summing over (k) corresponds to summing over the components of the mixture. Note that for the symmetric matrices $(\tilde{\Sigma}^k, (\Lambda\Sigma^k\Lambda'), \Sigma^\epsilon)$ the sum must only be taken over the unique entries to avoid double-counting. The restriction of summing only over $i \geq j$ achieves this.

The free parameters are $\Sigma^a, \Sigma^b, \mu^a, \mu^b$ and the non-normalised and non-zero entries of Λ. Intuitively, we are minimising the sum of the (squared) distances element-by-element. Note that this is linearly separable in components of the mixture, so adding additional components does not present a substantially greater computational challenge.

Given that the measures may not be taken in similar units, there is the risk of having very 'flat' sections of the criterion function, an issue that is discussed in the empirical section of this paper. To minimise the sum of squares function (23), I use the Limited-memory BFGS (L-BFGS) algorithm, a type of quasi-Newton method.[17]

3.4.3 Non-linear least squares

The previous two steps have provided an estimated distribution of latent factors. Next I draw a simulated dataset from this estimated mixture distribution. This is then treated as my 'observed' data, and that this is simulated data does not impact any remaining estimation. Using simulated data introduces noise, the impact of which vanishes as the size of the simulated dataset increases, so in general the size of the simulated dataset should be chosen to be as great as possible, given computational considerations.

To estimate the residuals $\hat{v}_{I,t}$, we must first estimate (9) by OLS. Equations (12) and (13) can then be estimated by non-linear least squares (NLS). A variety of optimisation techniques exist for doing so, of which I choose the Levenberg-Marquardt algorithm (Levenberg (1944) and Marquardt (1963)).[18] The Levenberg-Marquardt algorithm only finds local optima, so where concerns over multiple local optima apply a variety of starting values must be used to check whether this is indeed global.[19]

[17] The specific algorithm used is L-BFGS-B, which easily allows for box constraints. This is important in my case for constraining variances to be non-negative. See Broyden (1970) for an early description of the details. The algorithm was implemented using the 'stats' base R package.

[18] The algorithm was implemented using the 'minpack.lm' package in R.

[19] An alternative method that is less vulnerable to local optima is generalised simulated annealing, which has the disadvantage of being far more computationally challenging.

Although there exist analytical standard errors for the final NLS stage, these do not take into account the previous steps of estimation, in particular the simulation noise. Therefore in the empirical part of this paper I follow the previous literature and use bootstrap standard errors, taking into account the entire estimation process.

4 Empirical application: children in rural China

In this section I apply the three-step estimation procedure outlined in Section 3 to a very recent dataset, adding to the small and growing collection of papers that estimate these production functions in the existing literature. First I describe the data on which my analysis is based, before proceeding with simulation of factors and estimation of production functions. The majority of the analysis was performed in R version 3, with additional work in Stata 13 SE.[20]

4.1 Data description

In April 2013, Rural Education Action Programme (REAP) commenced a randomised controlled trial in Shaanxi Province, part of the Northwestern Region of the People's Republic of China. The first cohort recruited were 942 children of ages 6 to 12 months living in 351 villages. These children were followed up every 6 months until October 2014. In October 2013, an additional cohort of 860 children also aged 6 to 12 months was added. This second cohort was followed until April 2015. This gives four waves of data for each cohort. As demonstrated in Tables 2 and 3, the two cohorts share broadly similar observable characteristics. In our analysis we combine the two to give our sample, of which 52% of children are male. Although we do not know the total family size, 40% of children are not firstborn. The one-child policy does not apply to these households, as with many other rural families in China at this time. There are no two children from the same household in the sample.

[20] All code is my own and is available on request.

As reported in Luo et al. (2015), in the initial survey rounds children were found to have numerous health problems. Of the approximately 1800 children, 48.8% were found to be anaemic, having a deficiency of haemoglobin in their blood. Approximately a fifth of infants were significantly delayed in cognitive development and just under a third were significantly delayed in psychomotor development. This constitutes a 'severe' public health problem according to the World Health Organisation. Interestingly, the children exhibited low levels of stunting, underweight and wasting, suggesting the problem not to be lack of food, but poor quality of food.

The trial was to test the effectiveness of two interventions. As the interventions themselves are not the topic of interest of this paper, I only give brief details here and refer the reader to the REAP website for further information.[21]

The first intervention was a nutrition supplementation programme. Randomisation occurred at the village level, with villages randomised into three arms, two treatments and one control. In both treatment groups, children were given a nutrition supplement package consisting of iron, zinc, folic acid and various vitamins. In one of the two groups, this was re-enforced with reminders delivered via text messages.

The second intervention is based on the same lines as the famous Jamaica study, which delivered substantial benefits (Walker et al., 2005). This consisted of weekly home visits with the goal of improving parenting practices. The intervention started in November 2014, and was carried out on a subset of villages from the original sample. The allocation was approximately even across treatment and control groups from the nutrition intervention.

In my analysis I do not exploit the exogeneity of the variation induced by these interventions. I pool the treatment and control groups, and although the additional variance in investment is useful, the random assignment of treatment is not utilised. This is unlike the approach in Attanasio et al. (2015a), who estimate separate sets

[21] http://reap.fsi.stanford.edu/research/organization/6084/67097 gives an overview of the interventions. At the time of writing, there has been no formal evaluation of treatment effects published.

of production functions for the treatment and control groups. A third possibility would be to use the experimental variation as an instrument for investment. The discrete nature of the treatment variable made this infeasible, hence I have relied on the argument in Section 3 to use elements from the budget set as instruments.

An important consideration is whether it is valid to combine treatment and control groups. If the only effect of the treatment is to induce exogenous variation in investment, then there is no problem of pooling the groups. The exogenous variation is in fact beneficial for estimation of parameters. If in addition to influencing investment, treatment has an effect on the unobservable shock term, then investment is endogenous. Our control function approach is an appropriate remedy should this problem arise. Therefore I do not consider this to be a substantive criticism.

A more fundamental concern is that we may be think that a treatment alters the production function parameters themselves. Here there is previous evidence to draw on. In their study on Colombian data, Attanasio et al. (2015a) demonstrates that estimated parameters do not vary across treatment and control groups in a randomised control trial of a parenting skills intervention. Treatment effects are attributed fully to changes in inputs to the production function. This lends weight to the hypothesis that these parameters are indeed invariant to changes in policy. In principle, we can test the assumption that parameters are constant across treatment and control groups (or indeed any other sub-sample) in this paper , however due to sample size considerations this is not informative.

In each 6-monthly survey, information on a variety of measures related to the children and their families is recorded. These included two tests of child development, the Bayley-I test and the Ages and Stages Questionnaire. Haemoglobin levels, height and weight were recorded for children and parents. In addition to this, an extensive series of questions were asked on parenting behaviour and beliefs, in particular regarding nutrition of the child. Background information on income, education, and various wealth measures was also collected.

In the data cleaning stage, a number of observations were lost. Overall sample

attrition was low at around 8% over all four waves, however many children are missing in one or more wave, re-appearing in a later wave. An analysis of these children suggests that those who go missing from the sample for one or more waves tend to come from households with slightly higher incomes, and are marginally higher performing on cognitive tests. In Table 1 I present a comparison of children who are missing for one or more periods to those who are present throughout, using information from the first wave. The difference in household income is statistically significant. If I were to base my analysis on the whole sample whether fully-observed or not, this may be problematic. However, as explained shortly I restrict my sample to those who we observe in all waves. Therefore this difference will not bias my results.

Table 1: Attrition analysis

	Not missing (N=1215)		*Missing (N=580)*	
Statistic	Mean	St. Dev.	Mean	St. Dev.
PDI	89.69	17.86	90.95	16.35
MDI	96.67	16.96	97.09	16.81
Height (cm)	71.86	3.69	71.99	3.82
Weight (kg)	9.10	1.21	9.12	1.22
HH income (2013 Yuan)	28,855	26,705	32,930	28,092

(All variables observed at wave 1)
(Missing observations are those who are not present for one or more wave)

The raw data exhibited a small number of unrealistic outliers, likely due to data entry errors. I have excluded observations for which with one or more key variables have extremely unrealistic values. These include children with inconsistent ages given in different waves, as well as those of a height and weight that are not possible.

The particular EM algorithm I am using is not robust to observations with missing data, therefore I have had to exclude observations with any missing values for any of the variables that I use to infer factors.[22] This then includes the group

[22]The EM algorithm I use is included with the 'mixtools' package in R. The advantage of using this package is that it is computationally quick, however this comes at the cost of a lack of robustness to missing values.

of children that are missing for one or more waves. In total, this leaves a sample of 938 observations.

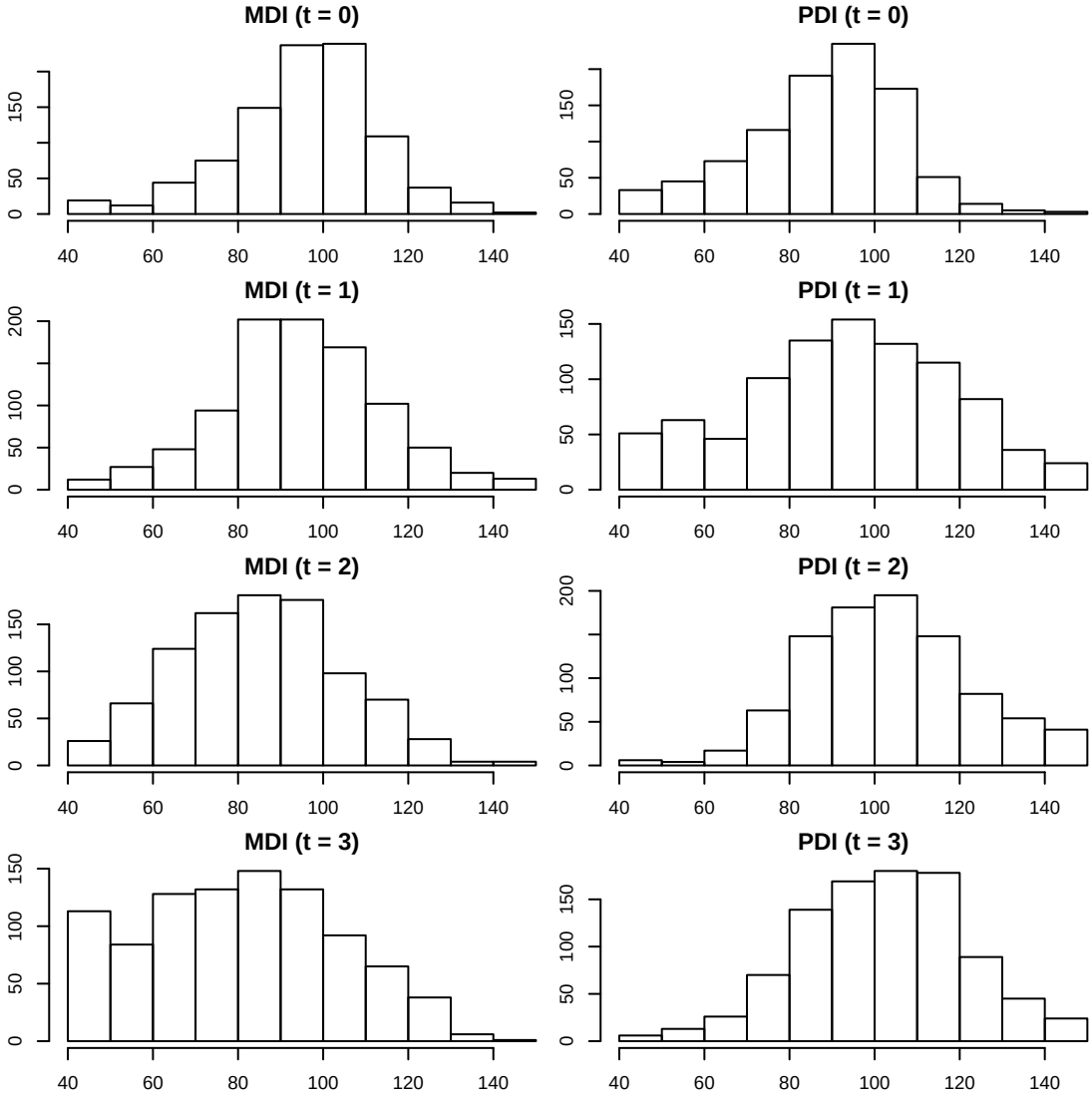
Table 2: Child-level summary statistics

	Cohort 1		*Cohort 2*	
Statistic	Mean	St. Dev.	Mean	St. Dev.
Wave 1 (6 - 12 months)				
Age in days	278.36	53.62	297.50	53.58
Height in cm	71.40	3.68	72.37	3.51
Weight in kg	8.98	1.15	9.22	1.20
Bayley MDI score	93.28	17.14	100.49	15.58
Bayley PDI score	83.64	17.35	95.79	15.74
Haemoglobin level	106.68	13.56	111.24	11.46
Wave 2 (12 - 18 months)				
Age in days	476.19	53.55	465.01	53.61
Height in cm	78.44	3.91	78.20	3.52
Weight in kg	10.26	1.20	10.28	1.22
Bayley MDI score	93.41	18.38	97.86	19.27
Bayley PDI score	95.58	23.28	93.44	25.71
Haemoglobin level	115.64	12.84	114.37	13.20
Wave 3 (18 - 24 months)				
Age in days	643.93	53.65	661.33	53.62
Height in cm	83.83	3.76	84.41	3.88
Weight in kg	11.30	1.25	11.54	1.44
Bayley MDI score	83.24	19.35	88.74	18.92
Bayley PDI score	98.42	17.09	110.70	19.85
Haemoglobin level	118.58	12.27	117.75	13.04
Wave 4 (24 - 30 months)				
Age in days	840.20	53.76	828.69	53.41
Height in cm	89.06	3.82	89.07	3.84
Weight in kg	12.59	1.39	12.73	1.63
Bayley MDI score	80.64	22.53	83.36	21.55
Bayley PDI score	100.47	18.65	105.26	19.26
Haemoglobin level	119.24	13.03	117.85	12.93

(Based on 938 observations)

Table 2 gives several useful summary statistics of the children, divided into the 6-monthly waves. The two cohorts are broadly similar on child characteristics, with one small exception discussed below.

Figure 1: Histograms of cognitive test scores



The Bayley-I test is a battery of developmental play tasks.[23] Raw scores are age-adjusted and combined to form two scores, a Mental Development Index (MDI) and a Psychomotor Development Index (PDI). Figure 1 presents histograms for each of the two measures, where the two cohorts are combined. Following the notation used throughout this section, $t = 0$ corresponds to wave 1, $t = 1$ to wave 2 and so on. One concern that is evident from the histograms is that they suffer a non-negligible truncation at the lower end of the distribution. Scores below 50 or above 150 are described as exceptional in the Bayley-I manual (Bayley, 1969), and are hence coded as 49 and 150 respectively. This is most evident for MDI at $t = 3$ and PDI at $t = 1$. In my analysis I keep this coded as though these are the true scores, however in future work using this data it would be fruitful to find an alternative solution to this issue. One possibility (which would be more complex when imbedded in a latent factor approach) would be to include the scores themselves along with dummy variables for those values coded as 49 and those coded as 150. Table 2 demonstrates that cohort 2 (the later cohort) tends to slightly outperform cohort 1 on cognitive tests.

Both Bayley scores are designed to have an expected mean of 100 and a standard deviation of 16, where the standardisation is based on a US dataset when the Bayley test was designed in the 1960s. We see then that both cohorts tend to underperform on the tests, with the exception of the later PDI scores for cohort 2. Scores below 80 correspond to a mild impairment, and scores below 70 a moderate or severe impairment. Bayley scales are age-adjusted, so do not necessarily increase over time.

The Ages and Stages Questionnaire (ASQ) is another screening tool used to assess child development and behaviour. Instead of testing the children directly, the ASQ uses parental observations. The questions involved are designed to indicate child development across numerous domains, and there is no standard way to create

[23] A recent comparison of the predictive power of the Bayley test and various other psychometric tests is provided in Richter et al. (2015).

a summary score. I am still however able to include several ASQ questions in my analysis to complement the Bayley scores.

Haemoglobin level is the number of grams of haemoglobin per litre of a subjects blood. A haemoglobin level of under 110 indicates anaemia, a condition where a lack of iron leads to a reduction in the number of red blood cells. As discussed in Luo et al. (2015), 49% of the sample in the first wave are classified as anaemic. Anaemia in early life is associated with many adverse outcomes (Grantham-McGregor and Ani, 2001), hence it is of interest when looking at child health.

Almost all children come from families in which the parents are married. The dataset contains information on the employment of the mother, and where the father is listed as the principal carer, the father also. The principal carer is listed as the mother for 48%, as the father for 44% and as a grandparent for the remaining children. The majority of mothers do not work and those who do, work in the town in which they live. Almost a third of fathers for whom we have data (those listed as principal carers) work in a different province or city.

Table 3: Household-level summary statistics

	Cohort 1		*Cohort 2*	
Statistic	Mean	St. Dev.	Mean	St. Dev.
Height of father in cm	170.12	5.55	170.09	5.71
Haemoglobin level of mother	156.88	5.64	157.28	6.09
Household income (2013 Yuan)	29,783	28,418	27,588	21,692
Durables index (0 - 9)	4.44	1.87	4.86	1.90
Principal carer education level (0 - 5)	1.87	0.80	1.89	0.87

(Based on 938 observations)
(All variables observed at wave 1)

Table 3 gives further information on the households. For reference, GDP per capita in China in 2013 was 25,362 Yuan.[24] Since many of the households in the dataset are likely to contain multiple working adults (including members of the wider family), the household income figure suggests a below-average per-capita income.

[24]World Bank databank, accessed 10/03/2016

The durables index is a sum of 0/1 indicators for ownership of 9 different types of durable goods, and can be seen as a proxy for wealth. The principal carer's education level is measured on a 0 to 5 scale, with 0 representing no schooling and 5 a college education and above. We can see that principal carers in this sample in general have low levels of education.

4.2 Choice of measures

The next step is to identify the measures I would like to include. Many of these have been presented in the summary statistics above. To choose measures, I used a combination of exploratory factor / correlation analysis and intuition, drawing on precedents in the literature where possible. As I assume a dedicated measurement system, ideal measures will be strongly correlated with a single unobserved factor of interest, and less correlated with other unobserved factors.

A further important consideration is that we hope to estimate a (continuous) mixture of normals distribution over these measures. As with previous papers in this literature, this presents a problem for measures that are discrete. The dataset here included many attractive measures that were either binary or took very few values, with no clear method of aggregation or of generating a continuous score.

In a Monte Carlo study, Attanasio et al. (2015b) found that including discrete measurements may not necessarily cause a large bias. I follow the literature here and include discrete variables without any smoothing adjustment, essentially treating them as continuous. The EM algorithm does not converge in the presence of a large number of discrete variables, which has limited the number of measurements I have been able to utilise. Nonetheless, for all but parental cognition I have identified a sufficient number of measures for each factor, some of which are discrete and some of which are (approximately) continuous.

Measures for cognitive ability come from the Bayley-I test and ASQ. Tables 4 and 5 present OLS regressions of cognitive measures on lagged health and cognitive measures and a set of household controls. As in other analysis in this section, wave 1

is labelled t = 0, wave 2 as t = 1 etcetera. These should not be interpreted as causal effects, but instead they serve to illustrate the partial correlations that are present in key measures in the sample. These estimates demonstrate the strong inertia of each cognitive measure. A comparison of tables 4 and 5 shows MDI in the second and third periods has strong predictive power for PDI in the next period, whereas only third-period PDI has predictive power for MDI in the following period.

There are also interactions between MDI, PDI and health measures that are difficult to explain, for example the significant negative effect of lagged weight on early MDI. This may in part be due to the influence of some unobserved factor that I have been unable to control for adequately, or more complex patterns of parental investment, which I do not directly control for here. An alternative explanation would be that our health measures are unsatisfactory and do not well-represent child health. This seems unlikely, particularly in the first wave since birthweight is a well-accepted measure of initial child health, and weight at age 6-12 months is very highly correlated with birthweight. We do however see height strongly predicting PDI in the earliest period, as would be expected.

We also see the model R^2 increasing over periods in both sets of regressions. This tells us that as children in the sample get older, more of the variance in cognitive test scores can be predicted by previous test performance, health variables and family background variables.

The dataset offers a number of objective and subjective health measures both for children and parents. One such measure I have used for children is a health index, related to 5 types of illness. This is calculated as 5 minus the sum of 0/1 indicators for whether or not a child has suffered from a specific type of illness in the last month. A higher number therefore reflects better health of the child. For parents I have opted only for the objective measures of haemoglobin level, height and weight.

In previous literature, it has been assumed that parental cognition can be represented by parental education and that this is measured without error. It would

Table 4: OLS Regression of MDI on lagged covariates

	Dependent variable:		
	MDI (t = 1)	MDI (t = 2)	MDI (t = 3)
	(1)	(2)	(3)
PDI (t - 1)	0.012	0.026	0.087**
	(0.041)	(0.026)	(0.035)
MDI (t - 1)	0.223***	0.286***	0.440***
	(0.041)	(0.034)	(0.035)
Height (t - 1)	0.247	−0.369	0.601***
	(0.218)	(0.211)	(0.228)
Weight (t - 1)	−1.437**	0.840	0.559
	(0.676)	(0.632)	(0.637)
Haemoglobin level (t - 1)	0.059	−0.068	0.073
	(0.046)	(0.046)	(0.051)
Family and Income controls	√	√	√
Constant	−0.997	75.009***	−67.561***
	(24.546)	(24.240)	(25.950)
Observations	938	938	938
R^2	0.097	0.132	0.253
Adjusted R^2	0.086	0.122	0.244
Residual Std. Error (df = 926)	18.016	18.096	19.258
F Statistic (df = 11; 926)	9.064***	12.783***	28.451***

Note: *p<0.1; **p<0.05; ***p<0.01
(Standard errors given in parenthesis)

Table 5: OLS Regressions of PDI on lagged covariates

	Dependent variable:		
	PDI (t = 1)	PDI (t = 2)	PDI (t = 3)
	(1)	(2)	(3)
PDI (t - 1)	0.136**	0.139***	0.340***
	(0.054)	(0.026)	(0.032)
MDI (t - 1)	0.039	0.163***	0.120***
	(0.056)	(0.034)	(0.032)
Height (t - 1)	1.003***	−0.170	−0.029
	(0.290)	(0.206)	(0.208)
Weight (t - 1)	0.692	−0.096	0.492
	(0.899)	(0.645)	(0.582)
Haemoglobin level (t - 1)	0.041	−0.113*	0.022
	(0.062)	(0.047)	(0.046)
Family and Income controls	√	√	√
Constant	−26.478	36.177	39.802*
	(32.628)	(24.740)	(23.722)
Observations	938	938	938
R^2	0.059	0.104	0.171
Adjusted R^2	0.048	0.094	0.162
Residual Std. Error (df = 926)	23.948	18.469	17.605
F Statistic (df = 11; 926)	5.249***	9.788***	17.413***

Note: *p<0.1; **p<0.05; ***p<0.01
(Standard errors given in parenthesis)

be preferable to model parental cognition as another unobserved factor informed by several measures, however the data requirements for this are clearly large and not met by my dataset. The variable I use for parental cognition is the educational level of the principal caregiver. A further complication is that in my dataset the principal carer is sometimes the mother, sometimes the father and on rare occasions a grandparent. As would be expected, mothers tend to be of lower education than fathers. With the above limitations in mind, I consider the education level of the principal carer a fair proxy for parental cognition.

Self-reported income and wealth is notoriously unreliable, and correlations in the data suggested measurement issues in our data. Therefore I treat wealth as a latent factor, complementing self-reported income with various alternative measures of wealth. One particularly attractive index is the sum over a set of durable goods in the home. In addition I have included information on house size and house price, which is often thought to be subject to less reporting error. I do not have information on home-ownership, however given extremely high national rates of ownership, it is unlikely that these are rental properties. In part due to the data that I have (many wealth measures are only measured in the baseline survey) I consider wealth fixed over time, only estimated a single wealth factor.

For investment measures, I drew on a number of sources. Ideally, these will reflect a number of different types of investment. For the later periods I used the number of food groups consumed by the child, information on supplements and a variable indicating whether the care-giver sang to the child, representing social interaction and stimulation. For the early period food groups are not appropriate, so I turn to alternative measures.

A potentially useful measure here is whether or not the child was given foreign-made prescription milk. In 2008 there was a major scandal involving Chinese milk and infant formula, which has led to a perception among many of it being unsafe.[25] Although any effect on child outcomes is unlikely to be direct, it is reasonable to

[25]Pei et al. (2011) give an overview of the scandal and its implications for the Chinese dairy sector.

think that whether a child is being given domestic or foreign prescription milk will be a good proxy for investment. Indeed, correlations between this, wealth, and other investment measures are high in the data, suggesting at least that there is meaningful variation in the measure. This will be discussed in more depth in the next section, in light of the estimated factor loadings.

The full set of measures used is presented along with factor loadings in Tables 6 and 7.

4.3 Simulation of factors

Once the measures have been selected, the next step in the estimation process is to fit the mixture of two normals distribution to the measures using the EM algorithm. I estimated a mixing parameter of 0.79, suggesting some deviation from normality. This is close to the mixing parameters of 0.73 and 0.65 found in Attanasio et al. (2015b).[26]

After the distribution of measures has been estimated, the next step is to apply the minimum distance step to find the distribution of factors. An issue is that the widely different scales on which variables are measured results in parts of the criterion function that are very flat-bottomed. This can lead to poor performance of the optimisation algorithm. To mitigate this problem I have normalised the variance of all measures to one before the minimum distance step.

Tables 6 and 7 present the estimated factor loadings for all 14 factors, corresponding to the λ parameters in Section 3. As can be seen, all but two factor loadings are positive where expected. The first anomalous loading is the child health index in the first period, which has a negative factor loading. It is not clear why this is, but my previous reduced-form analysis on the measures themselves indicates that the early health measures do exhibit correlations that are difficult to explain. The second surprising result is that the indicator for a child being given foreign prescription milk comes out as negative. Upon further investigation, there is indeed

[26] They estimate two mixture distributions, one for each cohort in their dataset.

a negative correlation between this measure and the number of supplements a child is given. Although each individual measure appears to be correlated with higher income and better child outcomes, the two measures are highly negatively correlated with one-another, explaining the opposite factor loadings.

As will be demonstrated in Table 8, neither of these anomalous measures (the first-period health index and the foreign milk indicator) is particularly informative to the factor of interest, so will be unlikely to influence results substantially.

A useful indicator to demonstrate the relationship between measures and factors is the signal-to-noise ratio. Letting $\tilde{\sigma}_i^2$ be the variance of measure i and σ_j^2 the variance of factor j, the signal to noise ratio for measure i and factor j is given by the ratio of the two $(\tilde{\sigma}_i^2/\sigma_j^2)$. Higher signal to noise ratios correspond to more informative measures. The full set of signal to noise ratios is reported in Table 8. For most factors there are a number of informative measures, with no single measure fully determining a factor. This demonstrates the usefulness of the latent factor approach, as factors are drawing on a variety of assigned measures.

For the health factor, it seems that height and weight provide most of the variation of the factor, with haemoglobin level and the child health index contributing relatively little. This observation is worthy of further discussion.

Correlational analysis and the discussion in Luo et al (2015) suggests that height and weight are not highly correlated with haemoglobin level in this dataset. Given this, if additional measures were available, it would be worth exploring the possibility of having two health factors, which has not been investigated in the literature. It could be possible for example to include a health factor representing longer-term health outcomes, less malleable in the short term, along with a factor representing shorter-term health outcomes that are rapidly influenced by other factors. Given the relatively small number of measures, this is not possible here. We thus maintain the single factor consisting of these three measures, complemented by the relatively uninformative health index. We must however be aware when interpreting the results that the health factor that this is primarily driven by height and weight.

Table 6: Estimated Factor Loadings

	Cognitive ability	Health	Investment
(t = 3)			
Bayley MDI score	1		
Bayley PDI score	0.764		
Does the child follow pointing?	0.204		
Child haemoglobin level		1	
Child weight (kg)		3.91	
Child health index		0.241	
Child height (cm)		3.838	
(t = 2)			
Bayley MDI score	1		
Bayley PDI score	0.735		
Does the child follow pointing?	0.051		
Child haemoglobin level		1	
Child weight (kg)		4.009	
Child health index		0.082	
Child height (cm)		3.353	
Number of Food Groups consumed			1
Number of supplements given to child			1.914
Do you sing songs to your child?			1.106
(t = 1)			
Bayley MDI score	1		
Bayley PDI score	0.714		
Child haemoglobin level		1	
Child weight (kg)		3.744	
Child health index		0.061	
Child height (cm)		3.179	
Number of Food Groups consumed			1
Number of supplements given to child			2.115
Do you sing songs to your child?			0.761
(t = 0)			
Bayley MDI score	1		
Bayley PDI score	0.774		
Child haemoglobin level		1	
Child weight (kg)		4.193	
Child health index		-0.402	
Child height (cm)		3.566	
Number of supplements given to child			1
Milk is foreign-produced			-0.276

Table 7: Estimated Factor Loadings

	Parental health	Wealth	Parental education
Mother haemoglobin level	1		
Father height (cm)	0.709		
Durables index		1	
House price (2013 Yuan)		0.912	
House size (meters squared)		1.434	
log(household income)		2.548	
Principal carer's education level			1

Parameters of the estimated distribution of factors are presented in Appendix B. From this distribution, I simulate the synthetic dataset of 1000 observations.[27] From this point onwards, I can treat this simulated data as my observed data, accounting for simulation noise by bootstrapping confidence intervals and standard errors.

Figures 2 and 3 show gaussian kernel density plots of the simulated factors. Here we can see notable departures from normality, demonstrating the benefit of assuming a mixture distribution instead of a single normal.

4.4 Results

In this section I present my production function estimates, taken over the simulated data. In Subsection 4.4.1 I ignore the endogeneity of investments. Then in Subsection 4.4.2 I estimate the investment equation and account for endogeneity via the inclusion of a control function in the production function.

4.4.1 Production function estimates with exogenous investment

Table 9 presents the estimated production function parameters for cognition, along with bootstrapped 95% confidence intervals.[28] The four waves of data allow us to

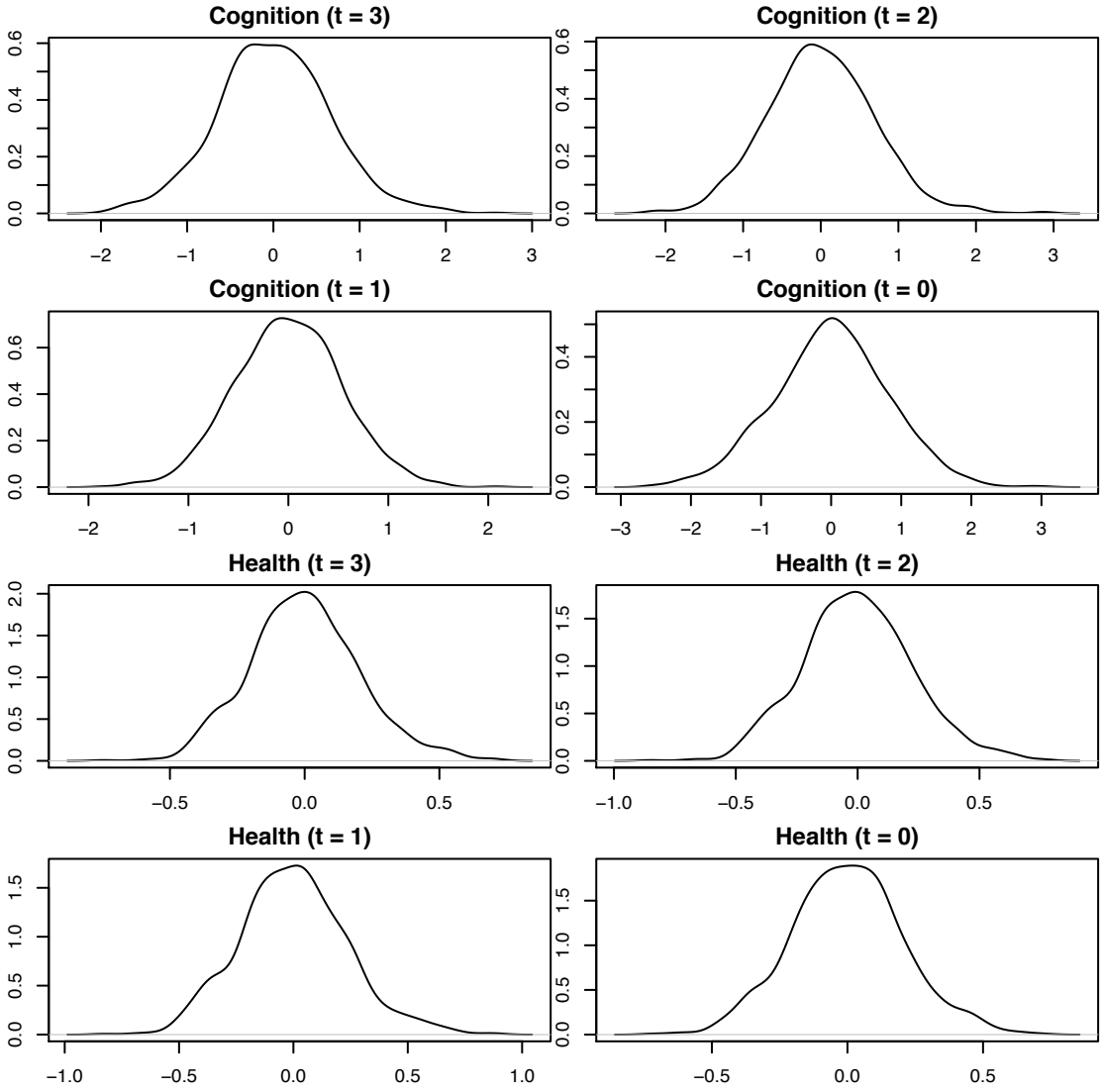
[27]There is no clear way to choose the number of simulated observations. I have chosen 1000 to be consistent with previous literature. Increasing this number does not affect results substantially, suggesting we can appeal to asymptotic results.

[28]In approximately 30% of bootstrap replications, either the EM algorithm or the algorithm used for non-linear least squares failed to converge. These replications are hence excluded from the confidence

Table 8: Signal to Noise ratios

Factor	Measure	Signal to noise ratio
Cognitive ability ($t = 3$)	Bayley MDI score	0.43
	Bayley PDI score	0.26
	Does the child follow pointing?	0.02
Health ($t = 3$)	Child haemoglobin level	0.04
	Child weight (kg)	0.52
	Child health index	< 0.01
	Child height (cm)	0.67
Cognitive ability ($t = 2$)	Bayley MDI score	0.48
	Bayley PDI score	0.26
	Does the child follow pointing?	< 0.01
Health ($t = 2$)	Child haemoglobin level	0.05
	Child weight (kg)	0.75
	Child health index	< 0.01
	Child height (cm)	0.61
Investment ($t = 2$)	Number of Food Groups consumed	0.11
	Number of supplements given to child	0.41
	Do you sing songs to your child?	0.13
Cognitive ability ($t = 1$)	Bayley MDI score	0.27
	Bayley PDI score	0.16
	Does the child communicate via language / gestures?	0.02
Health ($t = 1$)	Child haemoglobin level	0.05
	Child weight (kg)	0.75
	Child health index	< 0.01
	Child height (cm)	0.52
Investment ($t = 1$)	Number of Food Groups consumed	0.10
	Number of supplements given to child	0.44
	Do you sing songs to your child?	0.06
Cognitive ability ($t = 0$)	Bayley MDI score	0.66
	Bayley PDI score	0.41
Health ($t = 0$)	Child haemoglobin level	0.04
	Child weight (kg)	0.72
	Child health index	0.01
	Child height (cm)	0.57
Investment ($t = 0$)	Number of supplements given to child	0.41
	Milk is foreign-produced	0.03
Parental health	Mother haemoglobin level	0.23
	Father height (cm)	0.13
Wealth	Durables index	0.08
	House price (2013 Yuan)	0.07
	House size (meters squared)	0.11
	log(household income)	0.39

Figure 2: Densities of simulated factors



estimate three production functions. When interpreting my results, I refer to the earliest of the three as the first-period production function. Each row represents a different estimated parameter and each column a different production function. The estimating equation is re-stated in equation (24):

$$\theta_{c,t+1} = (\delta_{c,t}\theta_{c,t}^{\rho_t} + \delta_{h,t}\theta_{h,t}^{\rho_t} + \delta_{I,t}I_t^{\rho_t} + \delta_{ph,t}\theta_{ph}^{\rho_t} + \delta_{pc,t}\theta_{pc}^{\rho_t})^{1/\rho_t} exp(a_{c,t} + u_{c,t}) \quad (24)$$

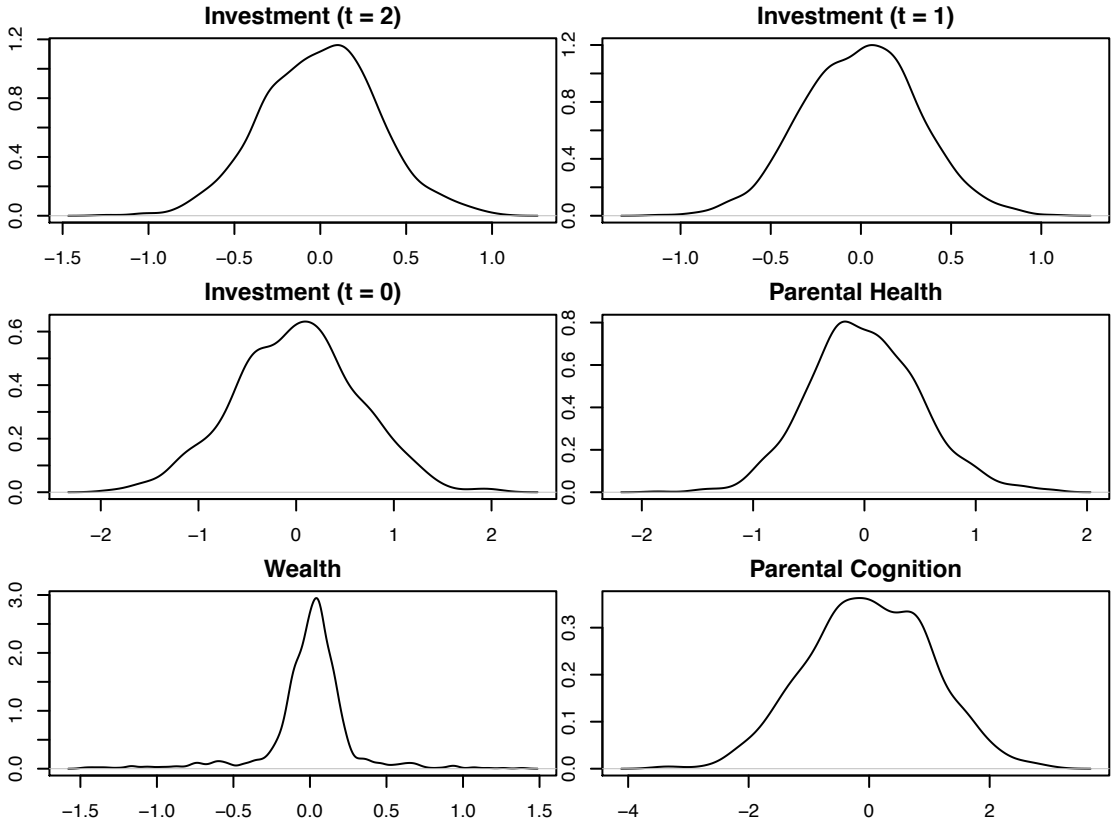
interval calculation.

The coefficient on TFP term $a_{c,t}$ is labelled 'TFP' in the estimates table. As would be expected given our normalisation of log-factors to have mean zero, this is not significantly different to zero in any period. I include it here for consistency with the theoretical model outlined in Section 3.

In Appendix B, I plot histograms of the bootstrap estimates for the parameters of the final-period cognitive ability production function. For several parameters, the distributions are not close to being normal. Hence I present confidence intervals in parentheses as opposed to standard errors.

While the confidence intervals are fairly large, in all periods, we cannot reject the hypothesis of a Cobb-Douglas ($\rho = 0$). As expected, there is a significant positive relationship between cognition across periods, which appears to stronger in the two later periods. If we accept the null of a Cobb-Douglas, we can interpret

Figure 3: Densities of simulated factors



these coefficients as elasticities. So in the first period, a 10% increase in lagged cognition leads to a 2.7% increase in current cognition. This rises to 9.4% in the second period and 8.7% in the final period.

There appears to be little effect of health on cognition. Lagged health in the first period is significant at the 10% level but not the 5% level. Investment is only significant and positive in the second-period production function. Parental health appears to have no effect throughout. Parental cognition on the other hand is significantly positive in the first period. This likely reflects some inherited characteristics that affect early cognition of the child, over and above those that are manifested in cognition at $t = 0$. Since cognition at $t = 0$ enters as an input for cognition at $t = 1$, we are controlling in later periods for this initial effect of parental cognition. There is no significant direct effect in later periods. We see a significant positive effect of investment in the second period, but not in any other. That the coefficients are not constant over time shows the importance of allowing for time-varying production function parameters.

Bringing these features together, Table 9 then suggests that early cognition is determined in part by parental cognition and potentially health, and as children grow older it is primarily previous cognition that predicts current cognition.

Table 10 presents the results for the health production function given in equation (25).

$$\theta_{h,t+1} = (\alpha_{c,t}\theta_{c,t}^{\xi_t} + \alpha_{h,t}\theta_{h,t}^{\xi_t} + \alpha_{I,t}I_t^{\xi_t} + \alpha_{ph,t}\theta_{ph}^{\xi_t} + \alpha_{pc,t}\theta_{pc}^{\xi_t})^{1/\xi_t} exp(a_{h,t} + u_{h,t}) \quad (25)$$

As in the cognition production function, in all cases we cannot reject a Cobb-Douglas function, although confidence intervals for ρ are large. We do see some effect of lagged cognition on health outcomes in the second period, but not in any other period. As expected, health in one period is a strong predictor of health in the next. It is worth remembering here that since all factors are normalised to

Table 9: Cognitive ability Production Function

	Cognition (t + 1)		
	(t + 1) = 1	(t + 1) = 2	(t + 1) = 3
ρ (t)	−0.020	0.713	0.097
	(−0.603, 0.317)	(−1.130, 0.898)	(−0.645, 0.634)
Cognition (t)	0.273	0.944	0.874
	(0.191, 0.473)	(0.437, 0.995)	(0.688, 1.014)
Health (t)	0.406	−0.249	−0.013
	(−0.038, 0.657)	(−0.526, 0.286)	(−0.325, 0.271)
Investment (t)	0.132	0.206	−0.123
	(−0.066, 0.178)	(0.067, 0.693)	(−0.309, 0.188)
Parental health	0.086	0.133	0.230
	(−0.070, 0.561)	(−0.232, 0.359)	(−0.047, 0.390)
Parental cognition	0.103	−0.033	0.032
	(0.017, 0.178)	(−0.095, 0.065)	(−0.018, 0.096)
TFP (t)	0.003	0.006	−0.005
	(−0.036, 0.071)	(−0.035, 0.048)	(−0.026, 0.030)

Note: 95 percent confidence intervals based on 500 bootstrap replications are given in parenthesis

have mean zero, the health factor does not on average increase over time, despite being driven by height and weight, which do of course increase with age. We appear to have unit-root behaviour in the health factor. As consistency of our estimator is based on $N \to \infty$ as opposed to $T \to \infty$, this is not a fundamental issue for consistency. However it could limit our ability to draw inference about long-run effects from these estimates.

Investment is not significantly different from zero in any period. Parental health has a positive impact in the first period and no impact in the later periods. Again, this first-period effect is likely capturing most of the heritability of height and weight. Note that we are still controlling for the child's initial height and weight, so this is likely demonstrating that children of larger parents tend to also grow quicker at the very earliest periods. Again, the change in estimated parameters over time demonstrates the need to allow for this flexibility. The estimated coefficient on parental cognition is not significantly different from zero in any period.

To summarise, the health production function results suggest a strong effect of previous health and some effect of parental health. Again there is little evidence of an impact of investment nor of any complementarities with cognition.

Across the two sets of estimated production functions, there seems to be a distinct difference between estimates in the first period and those in later periods. This lends support to the idea of critical periods, as discussed in earlier sections. As in both sets of production functions I cannot reject a Cobb-Douglas functional form, in Appendix B I re-estimate the model, imposing a Cobb-Douglas. My estimates are relatively unaffected by this restriction, demonstrating robustness and justifying my interpretation of the estimated coefficients as elasticities.

4.4.2 Endogenising investment

In this section, I address the endogeneity of investment by applying a control function approach. For this, I use the wealth factor Y_t, which was simulated with all other factors as explained in Subsection 4.3. As previously noted, I only simulate

Table 10: Health Production Function

	Health (t + 1)		
	(t + 1) = 1	(t + 1) = 2	(t + 1) = 3
ρ (t)	−0.279 (−0.726, 0.485)	0.279 (−0.505, 0.547)	−0.574 (−1.212, 1.492)
Cognition (t)	−0.013 (−0.039, 0.002)	0.048 (0.006, 0.053)	0.0004 (−0.012, 0.030)
Health (t)	0.967 (0.923, 1.004)	0.908 (0.873, 0.987)	0.981 (0.880, 1.017)
Investment (t)	−0.009 (−0.031, 0.006)	0.005 (−0.058, 0.071)	0.037 (−0.015, 0.115)
Parental health	0.061 (0.021, 0.122)	0.042 (−0.009, 0.084)	−0.018 (−0.054, 0.044)
Parental cognition	−0.005 (−0.020, 0.004)	−0.002 (−0.009, 0.009)	−0.0003 (−0.007, 0.014)
TFP (t)	−0.001 (−0.003, 0.002)	−0.003 (−0.005, 0.005)	0.001 (−0.005, 0.006)

Note: 95 percent confidence intervals based on 500 bootstrap replications are given in parenthesis

one wealth factor, so $Y_t = Y$ in every period.

The first step required is the estimation of the investment equation, approximated by a log-linear form, re-stated in equation (26) below:

$$ln(I_t) = \gamma_0 + \gamma_1 ln(Y_t) + \gamma_2 ln(\theta_{c,t}) + \gamma_3 ln(\theta_{h,t}) + \gamma_4 ln(\theta_{ph}) + \gamma_5 ln(\theta_{pc}) + v_{I,t} \quad (26)$$

Note that we require no condition on the relationship of any of the right-hand-side covariates with the error term $v_{I,t}$, as discussed in Section 3.

In Table 11 I present the estimated investment equation for each period. We see that investment is increasing in cognition and parental cognition in the two later periods, and in parental health in the first period. For my identification strategy, I require wealth to have a significant impact on investment. The p-value, inferred from the bootstrap replications, demonstrates that we can reject the null of a coefficient on wealth (γ_1) of zero at the 10% level in all three periods. My wealth factor predicts investment over and above all other input factors, satisfying the relevance condition. Given this, we proceed with the control function approach.

Table 12 presents the estimates for the cognition production function including the control function, as given in equation (27). Note once more that the shock term $u_{c,t}$ affects $\theta_{c,t+1}$. The timing of this is important for the interpretation of the control function estimates.

$$\theta_{c,t+1} = (\delta_{c,t}\theta_{c,t}^{\rho_t} + \delta_{h,t}\theta_{h,t}^{\rho_t} + \delta_{I,t}I_t^{\rho_t} + \delta_{ph,t}\theta_{ph}^{\rho_t} + \delta_{pc,t}\theta_{pc}^{\rho_t})^{1/\rho_t} exp(a_{c,t} + \hat{v}_{I,t} + u_{c,t}) \quad (27)$$

When the control function is included, we see a substantial increase in the coefficients on investment in the first two periods compared to the corresponding coefficients in Table 9. Investment is now positive and significant in both. This suggests that investment is compensatory, meaning that children with negative unobserved shocks to the cognition production process receive higher investment within that

Table 11: Investment Equation

	Investment (t)		
	t = 0	t = 1	t = 2
Cognition (t = 0)	−0.016 (−0.149, 0.149)		
Health (t = 0)	−0.176 (−0.908, 0.611)		
Cognition (t = 1)		0.137 (0.003, 0.222)	
Health (t = 1)		−0.113 (−0.298, 0.247)	
Cognition (t = 2)			0.193 (0.086, 0.310)
Health (t = 2)			−0.067 (−0.338, 0.329)
Parental health	0.435 (0.014, 0.744)	0.140 (−0.088, 0.237)	0.123 (−0.092, 0.265)
Parental cognition	0.074 (−0.017, 0.164)	0.048 (0.012, 0.094)	0.047 (0.001, 0.099)
Wealth	0.514 (0.047, 0.581)	0.140 (−0.011, 0.216)	0.082 (−0.051, 0.163)
p-value for wealth	0.008	0.054	0.09

Note: 95 percent confidence intervals based on 500 bootstrap replications are given in parenthesis. All variables in log form.

period. This can explain the insignificance of investment in Table 9. Since parents observe the negative shock and simultaneously increase investment in response, the effect of investment is biased downwards when we do not include the control function.

Investment is now the only significant predictor in the first period, bar lagged cognition. The p-values indicate that the coefficient on the control function residual is significant at the 5% level, in line with the previous argument of compensatory investment. In the final period, inclusion of the control function has less of an effect, indicating that endogeneity may be less important in this period.

Table 13 gives the estimates for health including the control function, equation (28).

$$\theta_{h,t+1} = (\alpha_{c,t}\theta_{c,t}^{\xi_t} + \alpha_{h,t}\theta_{h,t}^{\xi_t} + \alpha_{I,t}I_t^{\xi_t} + \alpha_{ph,t}\theta_{ph}^{\xi_t} + \alpha_{pc,t}\theta_{pc}^{\xi_t})^{1/\xi_t} exp(a_{h,t} + \hat{v}_{I,t} + u_{h,t}) \quad (28)$$

Introducing the control function here has an effect in the final period only. Investment in the final period becomes significantly positive and in both other periods remains not significantly different from zero. The significantly positive impact of investment is striking, given the previous discussion of how little the health factor seems to be affected by most inputs. It seems that despite the measures used for health being relatively 'long-term', investment can still have a short-term influence once endogeneity is properly accounted for. Again, the significant negative coefficient on the control function in the final period is consistent with compensatory behaviour.

All in all, the results for both cognition and health give several clear messages. Firstly, there is further evidence in support of a Cobb-Douglas functional form. They also suggest, consistent with previous literature, that investment is compensatory as opposed to complementary. Third, they do not indicate a significant role for complementarities between health and cognitive ability. Finally, we have shown that

Table 12: Cognitive ability Production Function with Control Function

	Cognition (t + 1)		
	(t + 1) = 1	(t + 1) = 2	(t + 1) = 3
ρ (t)	0.410 (−0.425, 0.400)	−0.026 (−0.772, 0.543)	0.090 (−0.502, 0.397)
Cognition (t)	0.231 (0.104, 0.432)	0.891 (0.303, 0.974)	0.894 (0.683, 1.164)
Health (t)	0.154 (−0.705, 0.534)	−0.490 (−1.275, 0.138)	0.041 (−0.331, 0.824)
Investment (t)	0.795 (0.028, 1.568)	0.568 (0.157, 1.662)	−0.203 (−0.988, 0.257)
Parental health	−0.188 (−0.417, 0.406)	0.120 (−0.221, 0.440)	0.227 (−0.124, 0.388)
Parental cognition	0.008 (−0.142, 0.132)	−0.089 (−0.185, 0.038)	0.040 (−0.004, 0.151)
TFP (t)	−0.033 (−0.043, 0.051)	−0.001 (−0.019, 0.019)	−0.005 (−0.016, 0.027)
Control function (t)	−0.719 (−1.542, 0.034)	−0.362 (−1.539, 0.152)	0.099 (−0.278, 1.040)
p-value for control function	0.034	0.044	0.116

Note: 95 percent confidence intervals based on 500 bootstrap replications are given in parenthesis

Table 13: Health Production Function with Control Function

	Health (t + 1)		
	(t + 1) = 1	(t + 1) = 2	(t + 1) = 3
ρ (t)	0.069	0.228	−0.216
	(−0.376, 0.347)	(−0.408, 0.499)	(−0.235, 0.500)
Cognition (t)	−0.008	0.032	−0.038
	(−0.040, 0.023)	(−0.020, 0.074)	(−0.123, 0.003)
Health (t)	1.006	0.868	0.871
	(0.925, 1.190)	(0.750, 1.118)	(0.622, 0.964)
Investment (t)	−0.102	0.067	0.193
	(−0.462, 0.042)	(−0.242, 0.276)	(0.035, 0.495)
Parental health	0.095	0.041	−0.012
	(0.017, 0.265)	(−0.006, 0.079)	(−0.054, 0.070)
Parental cognition	0.009	−0.008	−0.013
	(−0.012, 0.043)	(−0.023, 0.019)	(−0.050, 0.002)
TFP (t)	0.001	−0.002	−0.001
	(−0.003, 0.002)	(−0.004, 0.003)	(−0.002, 0.001)
Control function (t)	0.101	−0.085	−0.192
	(−0.047, 0.444)	(−0.276, 0.247)	(−0.502, −0.036)
p-value for control function	0.106	0.543	0.013

Note: 95 percent confidence intervals based on 500 bootstrap replications are given in parenthesis

investment has an effect on both health and cognitive ability, and that this effect varies over time. Given the short periodicity of the dataset at hand, it is natural to expect that any effects will be relatively small, so finding any effect of investment is a significant result, particularly on a health factor primarily determined by 'long-term' health measures.

4.5 Policy implications

The results presented in the previous subsection deliver several stylised facts. One advantage of taking an approach in which the parameters we identify are invariant to policy shifts is that it is possible to perform multiple policy experiments that would be infeasible to evaluate directly under a more reduced-form approach. Using the estimated model, it is possible to 'back out' the optimal path for investment. This is perhaps the most direct policy implication of the empirical results. To discuss the optimal path, we must first decide on an objective function that best fits a benevolent policy-maker's goals. Here I decide to use a simple objective function that can be thought of as the child's human capital in the final period. An attractive functional form for human capital will consist of health and cognitive ability, and will allow for some degree of complementarity between the two. A Cobb-Douglas function fulfils this, hence I define human capital in each period as:

$$HC_t = \theta_{c,t}^{\eta} \theta_{h,t}^{(1-\eta)} \tag{29}$$

As it is not clear a priori how to weight cognitive ability and health in overall human capital, varying the parameter η allows us to explore a number of different possible weightings. Given that both health and cognitive ability in early years have significant impacts on later life, any reasonable weighting would be fairly balanced, not putting all weight on one aspect of child development. Choices of η can be informed by other results in the literature, and inherently involve value judgements that I do not make here.

A further question is whether or not we include discounting. Given the short time frame (two years) involved in the data, I choose to abstract from any discount rate, which has the benefit of keeping it clear which factor is driving the result. This could be easily relaxed, and should the data be over a longer period a discount rate ought to be included.

In these experiments, I assume there is a total of one unit of investment that can be allocated in any way across the three periods. I abstract from any costs of investment. Under these simplifying assumptions, the optimal path will then depend on the marginal returns to investment in each period, which depend on all estimated coefficients as well as levels of inputs.

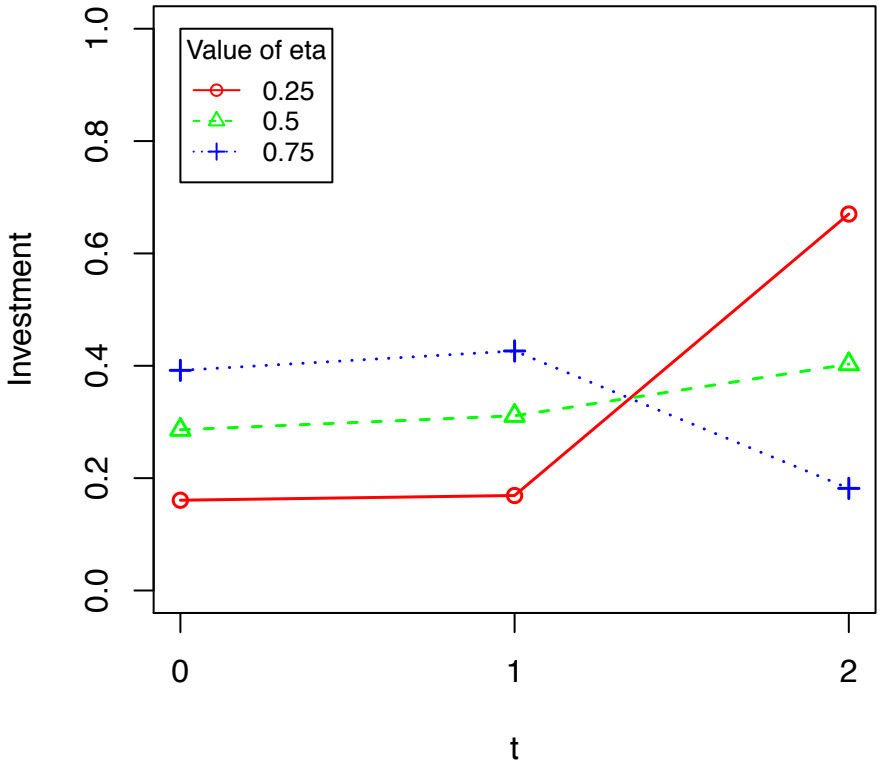
In all experiments, I assume a Cobb-Douglas structure. Hence I use the estimates laid out in Tables 22 and 23 in Appendix B. In addition, I set any estimated parameter that is not significant to zero. The full maximisation problem is presented in Appendix C.

Figure 4 presents the optimal investment path under three specifications, firstly when we weight cognitive ability more highly than health in the human capital function ($\eta = 0.75$), secondly when we weight the two equally ($\eta = 0.5$) and finally when health receives the greater weight ($\eta = 0.25$). All starting conditions including parental factors are fixed at their mean level in wave 0 of the sample. These are only point estimates, and do not reflect the uncertainty of the estimated production function parameters.

As can be seen in Figure 4, if cognitive skills and health are equally weighted, it is optimal to spread investment approximately evenly across all three periods. The higher the weight on health (lower η), the more investment is shifted towards the final period. A higher weight on cognition (higher η) leads to higher investment in the first two periods.

When focusing on cognition as the main outcome, we do indeed see that investment is best allocated to the earliest periods, where the returns are highest (see Table 22). When health is heavily weighted, we in fact almost get the opposite.

Figure 4: Optimal Investment Paths



This is driven by the insignificance of investment in all but the final period (see Table 23). With my estimated parameters, initial conditions do not substantially affect the optimal investment paths, so I do not present results where I vary these here.

Given the short time-span of this data, one ought to be cautious in making any direct policy recommendations based on the investment paths calculated above. However it is apparent that working in this framework allows policy recommendations in which the underlying assumptions are clear, and we can easily investigate implications of policy shifts. For example, it would be possible to calculate the implications for child development of a shift in investments from children at age 12-18 months to children age 0-6 months. If the data followed children for a longer period of time, we would be able to give a more detailed optimal path, which would be extremely useful from a policy perspective.

5 Simulation evidence

In this section I investigate how the distributional assumption used when accounting for measurement error affects the production function estimates. I entirely abstract from the endogeneity of investment issue in order to focus on the main results. The key question I seek to address is whether the assumption that measures (and hence log-factors) are distributed according to a mixture of normal distributions is consistent with the imposed CES functional form.

5.1 Benchmark simulations

In this subsection I abstract from measures, considering only the distribution of factors. Although the distributional assumptions are motivated by the unobservability of factors and hence the necessity to move from measures to factors, to illustrate the impact of the assumptions it is sufficient to use only the factors themselves. In Subsection 5.2 I demonstrate that introducing the measurement system does not qualitatively affect my results.

To simplify analysis further, assume that there are only three factors $\theta = (\theta_1, \theta_2, \theta_3)$. With three factors, we can form a production function with two inputs and one output. This can be extended easily to the case of many factors and is sufficient to illustrate the issue at hand. I also consider only a single, static production function to further isolate the potential source of bias.

5.1.1 Assuming a single normal distribution

To give an illustrative example of how a parametric assumption can influence our production function results, consider a hypothetical case in which we assume not that log-factors are distributed according to a mixture of two normals, but that they follow a single normal distribution:

$$ln(\theta) \sim N(\mu, \Sigma) \tag{30}$$

It is a well known feature of the normal distribution that any linear combination of normal variables is normally distributed itself. Imposing a Cobb-Douglas production function, in which $ln(\theta_1)$ is a weighted sum of $ln(\theta_2)$ and $ln(\theta_3)$ is then perfectly consistent with all three log-factors being normal.

The more general CES functional form, in which $ln(\theta_1)$ is a non-linear combination of $ln(\theta_2)$ and $ln(\theta_3)$ is then in general inconsistent with all three log-factors being normal, other than in the Cobb-Douglas special case. Given this argument, we would expect imposing a normal distribution on log-factors to bias estimated production function parameters towards a Cobb-Douglas.

Simulation evidence illustrates this point. The process I adopt in these simulations and that which all simulations in this section are based on is as follows:

1. Draw 1000 observations of (θ_2, θ_3) from an initial distribution[29]
2. Generate θ_1 using CES function (31) with a set of 'true' parameter values, adding a small amount of normally-distributed and independent noise u:[30]

$$\theta_1 = (\alpha {\theta_2}^{\rho} + \beta {\theta_3}^{\rho})^{(1/\rho)} exp(u) \tag{31}$$

$$u \sim N(0, \sigma^u)$$

3. Estimate the vector of means and covariance matrix of the simulated log-factors ($ln(\theta_1)$,$ln(\theta_2)$,$ln(\theta_3)$)
4. Draw 1000 observations from a log-normal distribution with the means and covariance matrix estimated in the previous step
5. Estimate a CES function using the observations drawn in step 4
6. Return to step 1 and repeat until the chosen number of replications is met

In step 1, the distribution of (θ_2, θ_3) is chosen to be a log-normal, with the parameters given below. The error variance is $\sigma^u = 0.02$.

[29] Choosing a greater number of observations does not affect the results and is computationally costly.

[30] This noise is not essential to the process and is kept small enough so that the parameters can be well recovered by non-linear least squares after step 2 (before the simulation process in step 4).

$$(ln(\theta_2), ln(\theta_3)) \sim N \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}$$

The results in Table 14 demonstrate that there is indeed a substantial bias towards zero in the estimated values of the ρ parameter, reinforcing my previous analytical argument. For ρ, which governs the elasticity of substitution, I choose a range of values from -1 to 1, representing a variety of realistic degrees of complementarity. Each row corresponds to a different 'true' ρ. For each parameter, I present the mean estimate over 50 runs and the mean squared error.

Table 14: Simulation evidence assuming a single normal

	$\hat{\rho}$	MSE($\hat{\rho}$)	$\hat{\alpha}$	MSE($\hat{\alpha}$)	$\hat{\beta}$	MSE($\hat{\beta}$)
ρ = -1	-0.0190	0.9628	0.9390	0.1929	0.0519	0.2008
ρ = -0.5	-0.0092	0.2410	0.8089	0.0955	0.1873	0.0978
$\rho = 0$	0.0021	0.00001	0.5016	0.00001	0.4984	0.00001
$\rho = 0.5$	0.0079	0.2423	0.1869	0.0980	0.8099	0.0961
$\rho = 1$	0.0175	0.9657	0.0516	0.2011	0.9400	0.1937

Based on 50 runs for each value of ρ with $\alpha = 0.5$ and $\beta = 0.5$

It is also apparent that there is a substantial bias in the estimates of α and β, which in these simulations are both fixed at 0.5. This benchmark is interesting given that Cobb-Douglas functional forms have been found in the empirical work across a variety of geographical locations, time periods and in contrasting economic situations.

If we assume that measures are normally distributed and maintain the assumed linear-log relationship between measurements and factors, the resulting factors are log-normally distributed also. Estimating a full CES function on factors derived from normally-distributed measures will similarly lead to the conclusion that the production function is Cobb-Douglas. If factors are indeed log-normal, in a sense there is no gain to allowing for the more flexible functional form of a CES, and it is more natural to directly estimate a Cobb-Douglas model. Given that a Cobb-

Douglas function can be estimated more simply than a full CES by linear regression on log-transformed variables as opposed to requiring non-linear optimisation, this is a non-trivial concern.

Many measures of interest are indeed approximately normally distributed, for example height, weight and test scores. Therefore one must be aware that when using many such measures we may be unlikely to find anything other than Cobb-Douglas. While this may not be a concern if we truly believe factors to be log-normal, it calls into question the value of the increased flexibility of the CES function in some circumstances.

5.1.2 Assuming a mixture distribution

In the approach outlined in Section 3 it is not assumed that factors follow a log-normal distribution, in part due to the inconsistency with the general CES form outlined above. Instead a mixture of log-normals is assumed, which maintains much of the analytical convenience of a normal distribution but allows for more flexibility.

As in the case of assuming a single normal, we hope to learn whether assuming a mixture of two normal distributions implies any constraints on the parameters of the CES function, in particular parameter ρ. If it is the case that this assumption leads to a bias towards Cobb-Douglas ($\rho = 0$), it may force us to look more carefully at the previous results in the literature finding a Cobb-Douglas form.

There exists an almost analogous result for mixtures of normals to that which is outlined for a single normal in the previous section. Any linear combination of a mixture of normal variables with identical mixing parameters is distributed according to a mixture of normals itself. This is a less commonly-known result than that for the single normal, for obvious reasons. Given this, I have included a simple proof in Appendix D. To apply this result, we must also observe that analogous to a single normal distribution, variables that jointly follow a mixture of normals do so marginally also.

In particular, this means that if the data ($ln(\theta_2)$,$ln(\theta_3)$) are assumed to be a

mixture of normals and we assume a Cobb-Douglas production function to generate $ln(\theta_1)$ such that $ln(\theta_1)$ is a linear combination of $ln(\theta_2)$ and $ln(\theta_3)$, this variable also follows a mixture of normals. If we were to fit a joint mixture of normals to all three variables, we would in theory be able to identify the true joint distribution. In the non-Cobb-Douglas CES function, the non-log-linearity induced by $\rho \neq 0$ will result in $ln(\theta_1)$ in general not being distributed according to a mixture of normals.

The above discussion is very similar to that which was presented in the case of a single normal. Unfortunately making clear statements such as 'the results are likely to be biased towards a Cobb-Douglas' is far less simple in this setting. This is because the mixing parameter τ is unknown, and must be estimated. In adjusting to fit all three factors, the EM algorithm uses both the mixing parameter and the parameters within each normal component in the mixture distribution. Therefore it is not clear a priori in which direction a bias would go.

Given the ambiguity of the direction of any bias when the assumption of a mixture of normals is made, the simulation approach adopted in the previous subsection can be usefully adapted. The full process here is as follows:

1. Draw 1000 observations of (θ_2, θ_3) from an initial distribution
2. Generate θ_1 using CES function (31) with a set of 'true' parameter values, adding a small amount of normally-distributed and independent noise u
3. Use the EM algorithm to fit a mixture of normals distribution over all three log-factors generated in steps 1 and 2
4. Draw 1000 observations from the estimated mixture of normals in step 3
5. Estimate a CES function using the observations drawn in step 4
6. Return to step 1 and repeat until the chosen number of replications is met

The key steps are 2 to 4, in which the process of fitting a mixture of normals will result in some degree of information loss on the true distribution of log-factors. With the simulations we hope to learn how this loss of information affects the estimation in step 5.

The first decision that must be made in the above process is which initial distribution is to be chosen for (θ_2, θ_3), the two input factors. I choose to assume that the logs of the two input factors follow a mixture of normals themselves. Any bias in the estimated parameters is hence caused by the approximation of $ln(\theta_1)$ as following a mixture of two normals, which in general will not hold precisely. Again, while this is a simplifying assumption it helps narrow down the area in which problems are occurring. I relax this assumption later, exploring other distributions for (θ_2, θ_3).

Secondly, we must decide which parameters to select in stage 2 when generating θ_1. As we are most interested in the substitution parameter ρ, I hold α and β fixed. Further simulations have demonstrated that the choice of these two parameters has no impact on the direction of any bias found.

Table 15: Benchmark simulation specifications

	$\hat{\rho}$	MSE($\hat{\rho}$)	$\hat{\alpha}$	MSE($\hat{\alpha}$)	$\hat{\beta}$	MSE($\hat{\beta}$)
Specification 1						
ρ = -1	-1.0092	0.0011	0.4976	0.0001	0.4971	0.0001
ρ = -0.5	-0.4987	0.00005	0.5002	0.000001	0.5006	0.00001
ρ = 0	0.0005	0.000004	0.4999	0.000000	0.5001	0.000000
ρ = 0.5	0.5011	0.0001	0.4994	0.00001	0.4999	0.000002
ρ = 1	1.0049	0.0008	0.4985	0.0001	0.4988	0.00005
Specification 2						
ρ = -1	-0.9046	0.0124	0.5246	0.0008	0.4577	0.0024
ρ = -0.5	-0.4766	0.0009	0.5148	0.0003	0.4803	0.0005
ρ = 0	0.0038	0.0005	0.5015	0.0001	0.4986	0.0001
ρ = 0.5	0.4733	0.0012	0.4808	0.0006	0.5145	0.0003
ρ = 1	0.9197	0.0096	0.4646	0.0019	0.5208	0.0006
Specification 3						
ρ = -1	-1.0265	0.0041	0.4904	0.0003	0.5098	0.0005
ρ = -0.5	-0.5099	0.0005	0.4961	0.0001	0.5043	0.0001
ρ = 0	-0.0010	0.00001	0.4994	0.000002	0.5006	0.000002
ρ = 0.5	0.5108	0.0005	0.5052	0.0001	0.4956	0.0001
ρ = 1	1.0346	0.0044	0.5157	0.0006	0.4881	0.0003

Based on 50 runs for each ρ value with true $\alpha = 0.5$ and $\beta = 0.5$
Parameters for each specification given in Appendix E

Table 15 presents the first set of simulation results. As can be seen, depending

on the parameters of the initial distribution of (θ_2, θ_3), there is either no (or minimal) bias (Specification 1), modest bias toward Cobb-Douglas (Specification 2), or perhaps more surprisingly a small bias away from Cobb-Douglas (Specification 3) for certain distributions.

We also see that the biases in the estimated values of α and β are trivial when compared to those that are found in table 14 (assuming a single normal distribution), and that the biases in the estimated values of ρ are also smaller. There is therefore a substantial gain in assuming a mixture of two normals instead of a single normal.

The chosen distribution for (θ_2, θ_3) in Specification 1 is given below:

$$(ln(\theta_2), ln(\theta_3)) \sim \tau f^a + (1-\tau) f^b$$

where:

$$\tau = 0.5, f^a = N\begin{bmatrix}0\\3\end{bmatrix}\begin{bmatrix}0.1 & 0.05\\0.05 & 0.1\end{bmatrix}, f^b = N\begin{bmatrix}5\\0\end{bmatrix}\begin{bmatrix}0.1 & 0.05\\0.05 & 0.1\end{bmatrix}$$

Parameters for all further specifications are given in Appendix E .For all specifications I use $\sigma^u = 0.02$.

The simulations provide support for my previous argument that when the production function is truly Cobb-Douglas, $ln(\theta_1)$ follows a mixture of two normals itself. In all specifications, when $\rho = 0$ the bias disappears and the 'true' production function parameters are recovered precisely.

As the initial distributions have been chosen somewhat arbitrarily in the above benchmark simulations, it is also useful to look at the presence and direction of biases in more realistic settings. Table 16 presents simulations in which the initial distribution of (θ_2, θ_3) is taken from a distribution estimated with real data. The estimated distribution used is that of cognition and investment reported in Attanasio et al. (2015a). As can be seen, this falls into the case in which the bias is away from Cobb-Douglas. This demonstrates that my results for Specification 3 in Table 15

are not an artefact of an unrealistic distribution of factors.

Table 16: Realistic simulation specifications

	$\hat{\rho}$	MSE($\hat{\rho}$)	$\hat{\alpha}$	MSE($\hat{\alpha}$)	$\hat{\beta}$	MSE($\hat{\beta}$)
Specification 4						
ρ = -1	-1.1293	0.0188	0.4927	0.0001	0.4907	0.0001
ρ = -0.5	-0.5291	0.0012	0.4989	0.00001	0.4991	0.00001
ρ = 0	0.0005	0.00001	0.4967	0.0001	0.5033	0.0001
ρ = 0.5	0.5375	0.0017	0.4992	0.00001	0.4986	0.00001
ρ = 1	1.1159	0.0152	0.4953	0.0001	0.4918	0.0001

Based on 50 runs for each ρ value with true $\alpha = 0.5$ and $\beta = 0.5$
Parameters for each specification given in Appendix E

One advantage of the simulation approach is that for (θ_2, θ_3), the true distribution is known. It is therefore fruitful to compare the true parameters of the distribution of these two variables with the parameters estimated by the EM algorithm when θ_1 is included. As expected, when the production function is Cobb-Douglas, the mixing parameter, as well as the means and variances of (θ_2, θ_3) are recovered perfectly. When it is not and the estimates are biased, there is distortion to the mixing parameter, means and variances, reflecting the fact that the three log-factors do not follow a joint mixture of normals.

The results in this subsection demonstrate that assuming all three log-factors follow a mixture of two normals in a setting where only two of them do follow a mixture of two normals may indeed introduce a bias. However it certainly does not constrain estimates of ρ to be close to zero, and may not even bias the estimate in this direction.

5.2 Including measures

Although the key questions of this section can be addressed without consideration of the measurement system, to get closer to the full estimation procedure I introduce measures.

The full simulation procedure for these results is very similar to the previous case, with two additional steps:

1. Draw 1000 observations of (θ_2, θ_3) from an initial distribution
2. Generate θ_1 using CES function (31) with a set of 'true' parameter values, adding a small amount of normally-distributed noise u
3. Using a set of factor loadings and adding normal measurement error, generate a set of measures[31]
4. Use the EM algorithm to fit a mixture of normals distribution over all measures generated in step 3
5. Use minimum distance to find the corresponding distribution of factors via the linking equations
6. Draw 1000 observations from the estimated mixture of normals in step 5
7. Estimate a CES function using the observations drawn in step 6
8. Return to step 1 and repeat until the selected number of replications is met

In Table 17 I present the results for the two specifications which I demonstrated lead to bias in the previous subsection. As the results demonstrate, once measures are introduced, the bias issue remains. Hence my abstraction from measures in the previous section is not misleading.

5.3 Assuming a mixture of three normals

In an extension to the standard approach, here I investigate how increasing the number of components in the mixture distribution affects parameter estimates in this controlled simulation environment.

As discussed in Section 3, we would hope that introducing further mixture components will result in a better approximation to the true distribution of factors.

[31] I generate three measures per factor in these simulations, approximately in line with the number of measures per factor used in the empirical application reported in Section 4. The factor loadings used are given in Appendix E.

Table 17: Simulation specifications including measures

	$\hat{\rho}$	MSE($\hat{\rho}$)	$\hat{\alpha}$	MSE($\hat{\alpha}$)	$\hat{\beta}$	MSE($\hat{\beta}$)
Specification 2						
ρ = -1	-0.9121	0.0117	0.5223	0.0008	0.4621	0.0021
ρ = -0.5	-0.4767	0.0011	0.5175	0.0005	0.4774	0.0009
$\rho = 0$	0.0089	0.0007	0.5039	0.0001	0.4962	0.0001
$\rho = 0.5$	0.4871	0.0013	0.4858	0.0005	0.5106	0.0003
$\rho = 1$	0.9374	0.0078	0.4729	0.0015	0.5157	0.0005
Specification 3						
ρ = -1	-1.0426	0.0061	0.4868	0.0004	0.5134	0.0007
ρ = -0.5	-0.5097	0.0005	0.4973	0.0002	0.5027	0.0003
$\rho = 0$	0.0002	0.0001	0.4998	0.00002	0.5003	0.00002
$\rho = 0.5$	0.5149	0.0007	0.5062	0.0001	0.4945	0.0001
$\rho = 1$	1.0358	0.0040	0.5107	0.0004	0.4883	0.0003

Based on 50 runs for each ρ value with true $\alpha = 0.5$ and $\beta = 0.5$
Parameters for each specification given in Appendix E

Given that the results of previous subsections show that the approximation using a mixture of two normals does indeed bias the estimated parameters, this could be a useful modification to the existing method. To avoid introducing additional complications, I return to the setting in which measures are not included, as in Subsection 5.1.

Table 18 compares estimates when a mixture of two normals is assumed to those when a mixture of three normals is assumed. The steps involved are just as described for the previous simulations without measures, but in the case of the mixture of three normals the EM algorithm is set to fit three components instead of two. As ρ is the main parameter of interest, I do not present the results for α and β. The first two columns show results when a mixture of two normals is assumed, and columns three and four for when a mixture of three normals is assumed.

Assuming a mixture of three normals results in lower standard deviations of estimates, which in almost all cases corresponds to a lower mean-squared error. In this sense, a mixture of three normals generally gives a better approximation. In these simulations however, it does not always result in a mean ρ estimate that is

closer to the true value, as is apparent from specification 8. In most cases, allowing for three mixture components does give an improvement both in terms of the mean parameter estimates and mean-squared error. Although I do not present the results here, assuming a mixture of three normals produces biases in the estimation of α and β that are no larger than those resulting from assuming a mixture of two normals.

In Specifications 5 and 8 $(ln(\theta_2), ln(\theta_3))$ follow a mixture of two normals, as was the case in Subsection 5.1.2. I have also included here two specifications where $(ln(\theta_2), ln(\theta_3))$ do not follow a mixture of two normals. In Specification 6 I instead choose for them each to follow a chi-square distribution. In Specification 7 I use a mixture of three normals, which provides another layer of complexity in the distribution of $(ln(\theta_2), ln(\theta_3))$. In both of these alternatives, assuming a mixture of two normals gives a large bias and assuming a mixture of three gives a distinct improvement.

5.4 Lessons for future applications

It may seem desirable to obtain rules that relate the true distribution of factors to the biases that emerge when they are approximated by a mixture of normals. However, after running a large number of specifications it is difficult to identify any clear pattern that would allow us to do so. Also, given that in empirical work the true distribution of factors is not known, it is difficult to see how such rules, if found would be generally applicable.

Therefore, the key message from these simulation results is that the loss of information when we approximate the distribution of factors does affect the estimated elasticity of substitution and that the direction of the bias varies.

In addition, I have demonstrated that allowing for more flexible distributions can indeed give a distinct improvement. The choice of a mixture of two normals as opposed to one is sensible given the discussion in Subsection 5.1.1. However it is not clear why a mixture of two normals is superior to a mixture of three, given that the latter will likely result in lower bias due to the increased flexibility. One important

Table 18: Comparison of mixture distribution assumption

	Two normals		Three normals	
	$\hat{\rho}$	MSE($\hat{\rho}$)	$\hat{\rho}$	MSE($\hat{\rho}$)
Specification 5				
ρ = -1	-0.9040	0.0125	-0.9701	0.0040
ρ = -0.5	-0.4758	0.0010	-0.4816	0.0008
ρ = 0	0.0035	0.0005	0.0017	0.0002
ρ = 0.5	0.4735	0.0011	0.4932	0.0008
ρ = 1	0.9191	0.0097	0.9606	0.0039
Specification 6 (chi square)				
ρ = -1	-1.4430	0.2262	-1.2404	0.0680
ρ = -0.5	-0.6870	0.0374	-0.5910	0.0087
ρ = 0	-0.00001	0.000000	-0.00005	0.000000
ρ = 0.5	0.6622	0.0284	0.5266	0.0270
ρ = 1	1.4663	0.2484	1.2021	0.1124
Specification 7 (mixture of three normals)				
ρ = -1	-0.7939	0.0513	-0.9490	0.0040
ρ = -0.5	-0.3480	0.0289	-0.4928	0.0002
ρ = 0	-0.000004	0.000000	0.0004	0.000000
ρ = 0.5	0.3416	0.0311	0.4916	0.0002
ρ = 1	0.7988	0.0448	0.9489	0.0038
Specification 8				
ρ = -1	-0.9860	0.0110	-1.0743	0.0065
ρ = -0.5	-0.4883	0.0007	-0.5051	0.0001
ρ = 0	0.0003	0.000000	0.0002	0.000001
ρ = 0.5	0.4933	0.0005	0.5036	0.0001
ρ = 1	0.9827	0.0116	1.0780	0.0075

Based on 50 runs for each ρ value with true $\alpha = 0.5$ and $\beta = 0.5$
Parameters for each specification given in Appendix E

consideration is computing power. It is true that for a high number of measures, the EM algorithm takes significantly longer to fit a mixture of three normals than a mixture of two. That being said, given the substantial reduction in approximation bias, the benefit is likely to outweigh the cost for many research projects. Extending in this direction and towards other, more flexible distributional assumptions seems like a natural next development for the three step estimator.

6 Conclusion

In this paper I have contributed to the child skill-formation literature in two clear ways.

Firstly, I have presented a set of empirical results using a new dataset, adding weight to the consistent findings of a Cobb-Douglas functional form in the literature. I am also able to present further evidence for compensatory investment behaviour, consistent with the current literature. My exercise in calculating the optimal investment path, although brief, illustrates a fundamental benefit of adopting the production function approach.

Secondly I have demonstrated how the approximation that factors and measures follow a mixture of two normal distributions impacts the production function estimates. The current assumptions in the literature can result in a bias either towards Cobb-Douglas or more surprisingly away from Cobb-Douglas. I have also shown that allowing for a more flexible distributional assumption can mitigate the problem.

One shortcoming of the empirical section of this paper is the lack of model validation. I do not have an easily-interpretable measure of in-sample fit. As explained in Keane (2010), this is an important aspect of assessing the worth of a model. The estimates do in many ways fit patterns that are known to hold in other settings, however out-of-sample prediction is complicated by the latent factor framework and by the specific and narrow time-span covered by the data. Once the dataset has

matured and there has been a full evaluation of the treatment effects of the interventions, it would be fruitful to compare these effects to the production function estimates here. As outlined in Todd and Wolpin (2003), these need not necessarily be the same, however a comparison of the two would be useful. The (criticism of) lack of model validation is a general feature of the education production function literature and is an area worthy of future investigation.

Given that the method used and investigated here is in its nascent stage, there are many additional avenues that would be useful to explore. There is no paper that I am aware of that looks at the full interplay between cognitive, non-cognitive and health outcomes. As results suggest that each of these impacts the other pair-wise, it would certainly be beneficial to do look at the full set of interactions between all three. The limiting factor up to this point has likely been data availability, however given the growth of studies such as that used in the empirical application here and the increased awareness of the importance of all three factors, this will likely change over the coming years.

A further direction is to explore the assumption that the technology follows a CES functional form. While this is a reasonable benchmark, more flexible functional forms do exist and are feasible to estimate computationally. A natural form to pursue would be the Translog function, which is a generalisation of the Cobb-Douglas that allows for more complex substitution patterns. Finally, the issue of discrete measurements is yet to be tackled fully in the literature. This type of data causes problems both theoretically and in the practical application of the EM algorithm. Finding a way to better deal with this problem would open up studies such as this one to inclusion of a wider set of discrete measures.

References

Almond, Douglas (2006), "Is the 1918 influenza pandemic over? long term effects of in utero influenza exposure in the post 1940 u.s. population." *Journal of Political Economy*, 114, 672–712.

Almond, Douglas and Janet Currie (2011), "Human capital development before age five." *Handbook of labor economics*, 4, 1315–1486.

Almond, Douglas, Lena Edlund, and Marten Palme (2009), "Chernobyl's Subclinical Legacy: Prenatal Exposure to Radioactive Fallout and School Outcomes in Sweden." *The Quarterly Journal of Economics*, 124, 1729–1772.

Attanasio, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina (2015a), "Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia." Working Papers 1046, Economic Growth Center, Yale University.

Attanasio, Orazio, Camila Fernández, Emla O. A. Fitzsimons, Sally M. Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina (2014), "Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in colombia: cluster randomized controlled trial." *BMJ*, 349.

Attanasio, Orazio, Costas Meghir, and Emily Nix (2015b), "Human Capital Development and Parental Investment in India." Cowles Foundation Discussion Papers 2026, Cowles Foundation for Research in Economics, Yale University.

Bailey, Drew, Greg Duncan, Candice Odgers, and Winnie Yu (2015), "Persistence and fadeout in the impacts of child and adolescent interventions." Technical report, LCC Working Paper Series.

Bayley, Nancy (1969), *Bayley scales of infant development: manual*. Psychological Corporation.

Becker, Gary S. (1964), *Human Capital*. Number 9780226041209 in University of Chicago Press Economics Books, University of Chicago Press.

Becker, Gary S. and Nigel Tomes (1979), "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility." *Journal of Political Economy*, 87, 1153–89.

Becker, Gary S. and Nigel Tomes (1986), "Human Capital and the Rise and Fall of Families." *Journal of Labor Economics*, 4, S1–39.

Behrman, Jere R. (1996), "The impact of health and nutrition on education." *The World Bank Research Observer*, 11, 23–37.

Behrman, Jere R. and Mark R. Rosenzweig (2004), "Returns to Birthweight." *The Review of Economics and Statistics*, 86, 586–601.

Bharadwaj, Prashant, Katrine Vellesen Løken, and Christopher Neilson (2013), "Early life health interventions and academic achievement." *American Economic Review*, 103, 1862–91.

Black, Sandra E., Aline Bütikofer, Paul J. Devereux, and Kjell G. Salvanes (2013), "This Is Only a Test? Long-Run Impacts of Prenatal Exposure to Radioactive Fallout." NBER Working Papers 18987, National Bureau of Economic Research, Inc.

Black, Sandra E., Paul Devereux, and Kjell G. Salvanes (2007), "From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes." Working Papers 200718, Geary Institute, University College Dublin.

Blömer, Johannes and Kathrin Bujna (2013), "Simple methods for initializing the em algorithm for gaussian mixture models." *Computing Research Repository*.

Broyden, C. G. (1970), "The convergence of a class of double-rank minimization algorithms 1. general considerations." *IMA Journal of Applied Mathematics*, 6, 76–90.

Carlsson, Magnus, Gordon B. Dahl, Björn Öckert, and Dan-Olof Rooth (2015), "The effect of schooling on cognitive skills." *Review of Economics and Statistics*, 97, 533–547.

Carneiro, Pedro and James J. Heckman (2003), "Human Capital Policy." NBER Working Papers 9495, National Bureau of Economic Research, Inc.

Cattan, Sarah, Emily Nix, and Sami Stouli (2016), "A three step estimator for nonlinear factor models." (work in progress).

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2011), "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *The Quarterly Journal of Economics*, 126, 1593–1660.

Coleman, James Samuel, Ernest Queener Campbell, and Carol J. Hobson (1966), *Equality of Educational Opportunity*. US Govt. Print. Off.

Cunha, Flavio and James J. Heckman (2007), "The Technology of Skill Formation." *American Economic Review*, 97, 31–47.

Cunha, Flavio and James J. Heckman (2008), "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation." *Journal of human resources*, 43, 738–782.

Cunha, Flavio, James J. Heckman, and Lance Lochner (2006), *Interpreting the Evidence on Life Cycle Skill Formation*, volume 1 of *Handbook of the Economics of Education*, chapter 12, 697–812. Elsevier.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach (2010), "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica*, 78, 883–931.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin (1977), "Maximum likelihood from incomplete data via the em algorithm." *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Figlio, David N., Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth (2013), "The Effects of Poor Neonatal Health on Children's Cognitive Development." NBER Working Papers 18846, National Bureau of Economic Research, Inc.

Fiorini, Mario and Michael P. Keane (2014), "How the allocation of childrens time affects cognitive and noncognitive development." *Journal of Labor Economics*, 32, 787–836.

Geweke, John F. and Michael P. Keane (1997), "Mixture of normals probit models." *Federal Reserve Bank of Minneapolis Staff Report*, 237.

Grantham-McGregor, Sally and Cornelius Ani (2001), "A review of studies on the effect of iron deficiency on cognitive development in children." *The Journal of nutrition*, 131, 649S–668S.

Hanushek, Eric A. (2003), "The failure of input-based schooling policies*." *The economic journal*, 113, F64–F98.

Heckman, James J. (2008), "Schools, Skills, and Synapses." IZA Discussion Papers 3515, Institute for the Study of Labor (IZA).

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith (1999), "The economics and econometrics of active labor market programs." *Handbook of labor economics*, 3, 1865–2097.

Heckman, James J. and Dimitriy V. Masterov (2007), "The productivity argument for investing in young children." *Applied Economic Perspectives and Policy*, 29, 446–493.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz (2010), "A New Cost-Benefit and Rate of Return Analysis for the Perry

Preschool Program: A Summary." IZA Policy Papers 17, Institute for the Study of Labor (IZA).

Heckman, James J., Rodrigo Pinto, and Peter Savelyev (2013), "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review*, 103, 2052–86.

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

Hoddinott, John, John A. Maluccio, Jere R. Behrman, Rafael Flores, and Reynaldo Martorell (2008), "Effect of a nutrition intervention during early childhood on economic productivity in guatemalan adults." *The lancet*, 371, 411–416.

Kautz, Tim, James J. Heckman, Ron Diris, Bas ter Weel, and Lex Borghans (2014), "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success." NBER Working Papers 20749, National Bureau of Economic Research, Inc.

Keane, Michael P. (2010), "Structural vs. atheoretic approaches to econometrics." *Journal of Econometrics*, 156, 3–20.

Knudsen, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff (2006), "Economic, neurobiological, and behavioral perspectives on building americas future workforce." *Proceedings of the National Academy of Sciences*, 103, 10155–10162.

Lemieux, Thomas (2003), "The "Mincer Equation" Thirty years after Schooling, Experience and Earnings." Working papers, University of California, Berkeley.

Levenberg, Kenneth (1944), "A method for the solution of certain non–linear problems in least squares."

Lucas, A., R. Morley, and T. J. Cole (1998), "Randomised trial of early diet in preterm babies and later intelligence quotient." *BMJ*, 317, 1481–1487.

Luo, Renfu, Yaojiang Shi, Huan Zhou, Ai Yue, Linxiu Zhang, Sean Sylvia, Alexis Medina, and Scott Rozelle (2015), "Micronutrient deficiencies and developmental delays among infants: evidence from a cross-sectional survey in rural china." *BMJ open*, 5, e008400.

Marquardt, Donald W. (1963), "An algorithm for least-squares estimation of nonlinear parameters." *Journal of the society for Industrial and Applied Mathematics*, 11, 431–441.

Mincer, Jacob A. (1974), *Schooling, Experience, and Earnings*. Number minc74-1 in NBER Books, National Bureau of Economic Research, Inc.

Newport, E. L. (2003), "Language development, critical periods in." In *Encyclopedia of Cognitive Science* (L. Nadel, ed.), Nature Publishing Group.

Norets, Andriy and Justinas Pelenis (2012), "Bayesian modeling of joint and conditional distributions." *Journal of Econometrics*, 168, 332–346.

Okun, Arthur M. (1975), *Equality and efficiency: The big tradeoff*. Brookings Institution Press.

Oreopoulos, Phil, Mark Stabile, Randy Walld, and Leslie Roos (2006), "Short, Medium, and Long Term Consequences of Poor Infant Health: An Analysis using Siblings and Twins." NBER Working Papers 11998, National Bureau of Economic Research, Inc.

Pearson, Karl (1894), "Contributions to the mathematical theory of evolution." *Philosophical Transactions of the Royal Society of London. A*, 185, 71–110.

Pei, Xiaofang, Annuradha Tandon, Anton Alldrick, Liana Giorgi, Wei Huang, and Ruijia Yang (2011), "The china melamine milk scandal and its implications for food safety regulation." *Food Policy*, 36, 412–420.

Richter, Linda, Musawenkosi Mabaso, and Celia Hsiao (2015), "Predictive power of psychometric assessments to identify young learners in need of early intervention: data from the birth to twenty plus cohort, south africa." *South African Journal of Psychology*.

Roberts, Brent W., Nathan R. Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R. Goldberg (2007), "The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes." *Perspectives on Psychological Science*, 2, 313–345.

Rubinstein, Yona and James J. Heckman (2001), "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review*, 91, 145–149.

Rust, John (2010), "Comments on: Structural vs. atheoretic approaches to econometrics by Michael Keane." *Journal of Econometrics*, 156, 21–24.

Rutter, Michael (2012), "Geneenvironment interdependence." *European Journal of Developmental Psychology*, 9, 391–412.

Schennach, Susanne M. (2004), "Estimation of nonlinear models with measurement error." *Econometrica*, 72, 33–75.

Teixeira, Pedro (2014), "Gary Beckers early work on human capital collaborations and distinctiveness." *IZA Journal of Labor Economics*, 3, 1–20.

Todd, Petra E. and Kenneth I. Wolpin (2003), "On the specification and estimation of the production function for cognitive achievement." *The Economic Journal*, 113, F3–F33.

Walker, Susan P., Susan M. Chang, Christine A. Powell, and Sally M. Grantham-McGregor (2005), "Effects of early childhood psychosocial stimulation and nutritional supplementation on cognition and education in growth-stunted jamaican children: prospective cohort study." *The Lancet*, 366, 1804–1807.

Wolpin, Kenneth I. (2013), *The Limits of Inference without Theory*, volume 1 of *MIT Press Books*. The MIT Press.

Wooldridge, Jeffrey M. (2015), "Control Function Methods in Applied Econometrics." *Journal of Human Resources*, 50, 420–445.

A The EM algorithm

The EM algorithm was formalised in Dempster et al. (1977), but had been applied to numerous special cases before this. The algorithm consists of an Expectation (E) step and a Maximisation (M) step within each iteration. In the E step, a likelihood function is created using the current estimates of parameters, and in the M step the likelihood function generated in the previous step is maximised, returning new parameters. Here I give a qualitative explanation of how the algorithm works in the context of Gaussian mixture models.

In the E step 'membership weights' are calculated for each data point and mixture component. These weights tell us the probability, given our current parameter set, that each data point came from each component and hence must sum to one for each data point across components. They do not reflect any 'mixing' property of each individual, as the component that each individual is drawn from is considered to be deterministic. Instead they reflect our Bayesian posterior on which is the relevant component, given the current parameter estimates and the observed data.

In the M step, the membership weights calculated in the E step are used to infer the mixture weights. With these mixture weights and the data, we can now find the mean and variance parameters that maximise the likelihood. These parameters are then returned to the E step. The combination of the two steps represents one iteration. Convergence is then based on the difference in log-likelihoods from one iteration to the next.

The EM algorithm is vulnerable to local optima. I found it to be common for the EM to pool all data points into one component, which in the production function context here results in a value of ρ close to zero. Given this, the choice of starting values for the EM algorithm is clearly important. After experimenting with several alternatives, I chose to use the K-means ++ algorithm. This provides estimates of cluster centres which can be used as starting values for the means of each component.[32] Starting values for variances of each component are the estimated

[32]Blömer and Bujna (2013) give a full comparison of the various options for initialising the EM algo-

sample variance of the data and starting values for the mixing parameter are chosen randomly.

B Further tables and figures

Table 19 presents the estimated means and variances for the 14 factors along with the mixing parameter across the two components. The full estimated distribution (not presented here) additionally includes all covariances between the 14 factors within each component.

Table 19: Estimated distribution of Factors

	Component A		*Component B*	
Mixing parameter	0.79		0.29	
	Mean	Variance	Mean	Variance
(t = 3)				
Cognitive ability	-0.0514	0.4031	0.2100	0.4753
Health	-0.0333	0.0337	0.1313	0.0515
(t = 2)				
Cognitive ability	-0.0300	0.4208	0.0911	0.6093
Health	-0.0373	0.0442	0.1432	0.0550
Investment	-0.0222	0.1096	0.0441	0.1005
(t = 1)				
Cognitive ability	0.0030	0.2767	-0.0224	0.4083
Health	-0.0352	0.0457	0.1271	0.0692
Investment	-0.0200	0.1033	0.0317	0.0880
(t = 0)				
Cognitive ability	-0.0155	0.6727	0.0564	0.7118
Health	-0.0249	0.0392	0.0985	0.0423
Investment	-0.0058	0.3968	0.0198	0.4867
Parental cognition	-0.0953	0.1999	0.2864	0.3162
Wealth	0.0140	0.0112	-0.0498	0.3049
Parental education	0.0331	0.9915	-0.0678	1.1292

In Tables 20, 21, 22 and 23 I present estimation results in which the Cobb-Douglas structure is imposed. In all specifications the estimated TFP term is pre-

rithm and demonstrate the effectiveness of K-means ++.

cisely zero (by construction), so is omitted.

Table 20: Cognitive ability Production Function

	Cognition (t + 1)		
	(t + 1) = 1	(t + 1) = 2	(t + 1) = 3
Cognition (t)	0.292 (0.195, 0.467)	0.964 (0.456, 1.040)	0.890 (0.685, 1.015)
Health (t)	0.319 (−0.067, 0.652)	−0.300 (−0.513, 0.306)	−0.014 (−0.388, 0.249)
Investment (t)	0.106 (−0.071, 0.174)	0.258 (0.062, 0.645)	−0.132 (−0.265, 0.219)
Parental health	0.208 (−0.079, 0.567)	0.099 (−0.295, 0.392)	0.241 (−0.055, 0.484)
Parental cognition	0.074 (0.025, 0.178)	−0.022 (−0.096, 0.077)	0.015 (−0.031, 0.090)

Note: 95 percent confidence intervals based on 500 bootstrap replications are given in parenthesis

Table 21: Health Production Function

	Health (t + 1)		
	(t + 1) = 1	(t + 1) = 2	(t + 1) = 3
Cognition (t)	−0.013 (−0.038, 0.001)	0.039 (0.011, 0.052)	0.005 (−0.012, 0.030)
Health (t)	0.962 (0.921, 1.003)	0.916 (0.852, 0.979)	0.959 (0.894, 1.014)
Investment (t)	−0.011 (−0.029, 0.008)	0.003 (−0.058, 0.086)	0.038 (−0.018, 0.105)
Parental health	0.068 (0.021, 0.124)	0.039 (−0.014, 0.077)	−0.005 (−0.059, 0.044)
Parental cognition	−0.006 (−0.022, 0.003)	0.002 (−0.009, 0.010)	0.002 (−0.008, 0.014)

Note: 95 percent confidence intervals based on 500 bootstrap replications are given in parenthesis

Table 22: Cognitive ability Production Function with Control Function

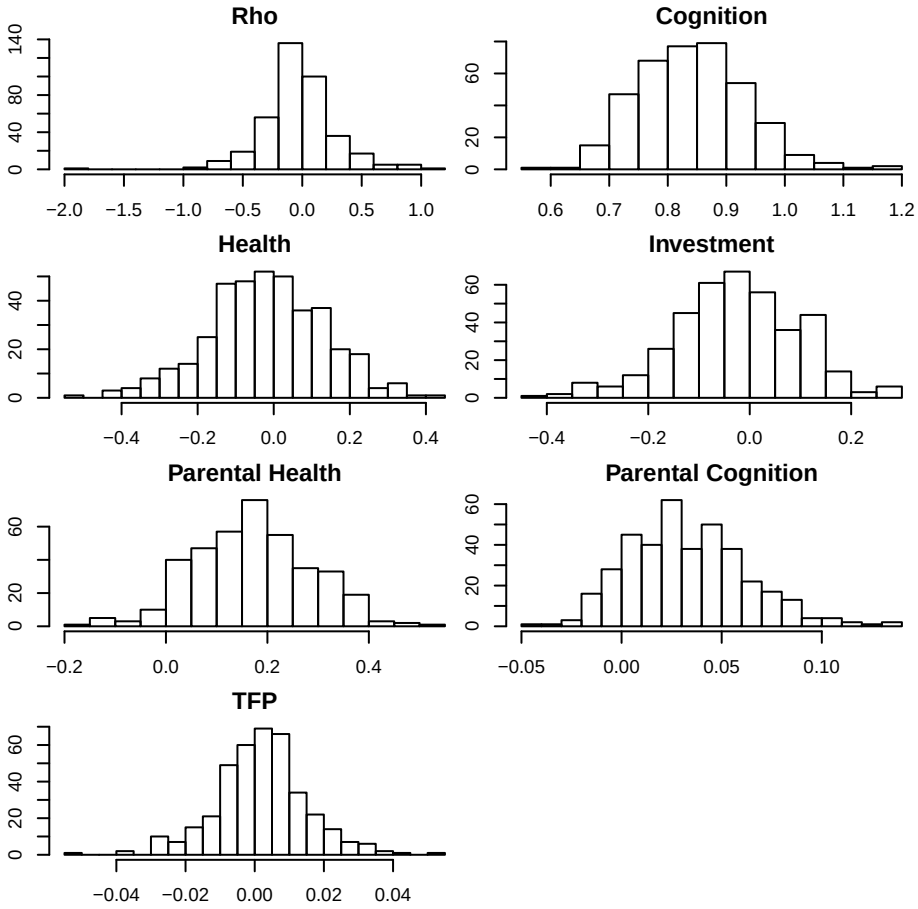
	Cognition (t + 1)		
	(t + 1) = 1	(t + 1) = 2	(t + 1) = 3
Cognition (t)	0.286 (0.077, 0.447)	0.900 (0.307, 0.953)	0.931 (0.690, 1.169)
Health (t)	0.086 (−0.737, 0.576)	−0.514 (−1.174, 0.141)	0.142 (−0.313, 0.720)
Investment (t)	0.552 (−0.045, 1.752)	0.543 (0.292, 1.548)	−0.320 (−0.916, 0.181)
Parental health	0.056 (−0.464, 0.436)	0.122 (−0.295, 0.469)	0.215 (−0.114, 0.466)
Parental cognition	0.020 (−0.167, 0.145)	−0.051 (−0.175, 0.034)	0.032 (−0.016, 0.164)
Control function (t)	−0.486 (−1.724, 0.111)	−0.381 (−1.460, 0.074)	0.235 (−0.231, 1.059)
p-value for Control Function	0.053	0.042	0.102
Note:	95 percent confidence intervals based on 500 bootstrap replications are given in parenthesis		

Table 23: Health Production Function with Control Function

	Health (t + 1)		
	(t + 1) = 1	(t + 1) = 2	(t + 1) = 3
Cognition (t)	−0.012 (−0.037, 0.034)	0.024 (−0.012, 0.076)	−0.030 (−0.116, 0.009)
Health (t)	1.002 (0.927, 1.227)	0.866 (0.731, 1.101)	0.825 (0.622, 0.971)
Investment (t)	−0.088 (−0.459, 0.052)	0.070 (−0.235, 0.257)	0.200 (0.008, 0.510)
Parental health	0.095 (0.015, 0.266)	0.045 (−0.013, 0.076)	0.018 (−0.061, 0.076)
Parental cognition	0.003 (−0.015, 0.050)	−0.005 (−0.019, 0.020)	−0.013 (−0.042, 0.002)
Control function (t)	0.084 (−0.054, 0.438)	−0.090 (−0.270, 0.235)	−0.202 (−0.520, −0.022)
p-value for Control Function	0.094	0.388	0.018

Note: 95 percent confidence intervals based on 500 bootstrap replications are given in parenthesis

Figure 5: Histograms of bootstrap coefficient estimates



In Figure 5 I present histograms of the estimated parameters across all 500 bootstrap replications for the final-period cognitive ability production function presented in Table 9 (omitting those where convergence failed). These demonstrate approximate normality with some deviations, for example the distribution of the estimates of ρ and the coefficient on cognition ($\delta_{c,t}$).

C Optimal investment

The full problem of the optimal level of investment is:

$$\max_{I_0,I_1,I_2} HC_3 = \theta_{c,3}^{\eta}\theta_{h,3}^{(1-\eta)}$$

$$\text{s.t.}$$

$$\sum_{t=0,1,2} I_t = 1$$

$$\theta_{c,t+1} = f_t^c(\theta_{c,t}, \theta_{h,t}, I_t, \theta_{ph}, \theta_{pc}) \qquad t = 0, 1, 2$$

$$\theta_{h,t+1} = f_t^h(\theta_{c,t}, \theta_{h,t}, I_t, \theta_{ph}, \theta_{pc}) \qquad t = 0, 1, 2$$

where f_t^c and f_t^h are the estimated production functions

The constraints can be sequentially substituted to form a non-linear maximisation problem subject to a linear constraint. I solve this by using a variant of the simulated annealing algorithm.

D Proof of mixture result

The probability density function for a mixture of normals is:

$$p(x|\tau, \mu_k, \Sigma_k) = \sum_k \tau_k \phi(x|\mu_k, \Sigma_k)$$

where $\phi(x|\mu_k, \Sigma_k)$ is the normal probability density function. Now let us consider summing h mixtures of normals, with the same mixing probability τ.

$$\begin{aligned}\tilde{x} := \sum_h x_h &= \sum_h \sum_k \tau_k \phi(x_h|\mu_{k,h}, \Sigma_{k,h}) \\ &= \sum_k \tau_k \sum_h \phi(x_h|\mu_{k,h}, \Sigma_{k,h}) \\ &= \sum_k \tau_k \sum_h \phi(x_h|\mu_{k,h}, \Sigma_{k,h}) \\ &= \sum_k \tau_k \phi(\tilde{x}|\tilde{\mu}_k, \tilde{\Sigma}_k)\end{aligned}$$

The final equality follows from the result that any sum of normally distributed variables is also normally distributed itself. Parameters $(\tilde{\mu}_k, \tilde{\Sigma}_k)$ are functions of the parameters of the initial individual component distribution.

E Simulation specifications

E.1 Initial factor draws

In all specifications except 6 and 7:

$$(ln(\theta_2), ln(\theta_3)) \sim \tau f^a + (1-\tau) f^b$$

where parameters in each specification are:

Specification 1

$$\tau = 0.5, f^a = N \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}, f^b = N \begin{bmatrix} 5 \\ 0 \end{bmatrix} \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}$$

Specification 2

$$\tau = 0.5, f^a = N \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0.1 & 0.06 \\ 0.06 & 0.08 \end{bmatrix}, f^b = N \begin{bmatrix} 0 \\ 2 \end{bmatrix} \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}$$

Specification 3

$$\tau = 0.2, f^a = N \begin{bmatrix} 2 \\ 5 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.6 \end{bmatrix}, f^b = N \begin{bmatrix} 3 \\ 4 \end{bmatrix} \begin{bmatrix} 0.8 & -0.03 \\ -0.03 & 0.1 \end{bmatrix}$$

Specification 4

$$\tau = 0.84, f^a = N \begin{bmatrix} 0.057 \\ 0.088 \end{bmatrix} \begin{bmatrix} 0.636 & 0.170 \\ 0.170 & 0.800 \end{bmatrix}, f^b = N \begin{bmatrix} -0.295 \\ -0.455 \end{bmatrix} \begin{bmatrix} 0.359 & 0.165 \\ 0.165 & 0.923 \end{bmatrix}$$

Specification 5

$$\tau = 0.5, f^a = N \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0.5 & 0.05 \\ 0.05 & 0.5 \end{bmatrix}, f^b = N \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0.3 & -0.1 \\ -0.1 & 0.5 \end{bmatrix}$$

Specification 6 - Chi Square

$$ln(\theta_2) \sim \chi^2(5)$$

$$ln(\theta_3) \sim \chi^2(3)$$

Specification 7 - Mixture of 3 normals

$$(ln(\theta_2), ln(\theta_3)) \sim \tau_1 f^a + \tau_2 f^b + \tau_3 f^c$$

$$\tau_1 = 0.5$$

$$\tau_2 = 0.3$$

$$\tau_3 = 0.2$$

$$f^a = N \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}, f^b = N \begin{bmatrix} 5 \\ 0 \end{bmatrix} \begin{bmatrix} 0.8 & 0.15 \\ 0.15 & 2 \end{bmatrix}, f^c = N \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0.05 \\ 0.05 & 0.5 \end{bmatrix}$$

Specification 8

$$\tau = 0.7, f^a = N \begin{bmatrix} 1 \\ 5 \end{bmatrix} \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}, f^b = N \begin{bmatrix} 4 \\ 2 \end{bmatrix} \begin{bmatrix} 0.3 & -0.1 \\ -0.1 & 0.5 \end{bmatrix}$$

E.2 Measurement system

In the simulations where measures are included, the factor loading matrix is as follows:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1.8 & 0 & 0 \\ 0.4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.3 & 0 \\ 0 & 3.1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1.5 \\ 0 & 0 & 0.8 \end{bmatrix}$$

For each factor, I attributed one measure with low measurement error, one with medium and one with high measurement error, to replicate a 'realistic' set of measures of varying degrees of informativeness about each factor.