Parental Investment, School Choice, and the Persistent Benefits of Intervention in Early Childhood

Lei Wang, Yiwei Qian, Nele Warrinnier, Orazio Attanasio, Scott Rozelle, and Sean Sylvia

Working Paper No. 931 October 2021

ISSN 1473-0278

School of Economics and Finance

Queen Mary University of London

Parental Investment, School Choice, and the Persistent Benefits of Intervention in Early Childhood

Lei Wang, Yiwei Qian, Nele Warrinnier,

Orazio Attanasio, Scott Rozelle, and Sean Sylvia*

October 2021

Abstract

We present evidence from a randomized experiment testing the impacts of a six-month early childhood home-visiting program on child outcomes at school entry. Two and a half years after completion of the program, we find persistent effects on child working memory - a key skill of executive functioning that plays a central role in children's development of cognitive and socio-emotional skills. We also find that the program had persistent effects on parental time investments and preschool enrolment decisions. Children were enrolled earlier and in higher quality preschools, the latter reflecting a shift in preferences over preschool attributes toward quality. Our findings imply an important role for the availability of high-quality subsequent schooling in sustaining the impacts of early intervention programs.

JEL Classification: J13, I21, I28, H11

Keywords: Early Childhood Development, Parenting, China, Poverty

*Wang: Shaanxi Normal University (email:wangleiml@snnu.edu.cn); Qian: Stanford University (email: yiweiqia@stanford.edu); Warrinnier: Queen Mary University of London and LICOS, KU Leuven(email: n.warrinnier@qmul.ac.uk); Attanasio: Yale University, the Institute of Fiscal Studies (IFS), and the Centre for Experimental Research on Fairness, Inequality and Rationality (FAIR) at the NHH Norwegian School of Economics (email: orazio.attanasio@yale.edu); Rozelle: Stanford University (email: rozelle@stanford.edu); Sylvia: University of North Carolina at Chapel Hill (email: sean.sylvia@unc.edu). We thank seminar participants at North Carolina State University, Queen Mary University of London, KULeuven, the annual conference of the Society of Labour Economists, the annual KDIS-3ie-ADB-ADBI Conference on Impact Evaluation at the Asian Development Bank, and the annual CEPR Women in Economics Workshop in Micro- & Applied Microeconomics. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 846621. We thank the International Initiative for Impact Evaluation (3ie) for financial support.

1 Introduction

Motivated by a growing body of research implying large private and social returns to investments in children at an early age, there has been interest among policy makers in cost-effective programs to improve early-life environments. The returns of these programs are likely to be particularly high in low and middle-income countries where it has been estimated that 43% of children under 5 are at risk of not reaching their full developmental potential (Black et al., 2017). Although several correlates of poverty contribute to this risk, a major factor appears to be that children often lack sufficiently stimulating environments during critical periods of brain development. Due to high levels of neuroplasticity, stimulation during this period is thought to be critical for longer term cognitive and psychosocial development (Knudsen et al., 2006).

One approach that has proven effective in improving child development outcomes is interventions that support caregivers to engage in stimulating activities with their children. Evidence from multiple randomized evaluations across several countries shows that parenting interventions can significantly improve the cognitive, language and socio-emotional skills of young children. Large-scale parenting interventions in Colombia (Attanasio et al., 2014), Pakistan (Yousafzai et al., 2014) and China (Sylvia et al., 2020; Heckman et al., 2020) have yielded sizeable effects on child outcomes in the short run. Although evidence on longer-run effects comes mostly from smaller-scale efficacy trials, studies have found promising results on a wide range of adult health, and labour market and social outcomes: increased college attendance, employment, and earnings and reductions in teen pregnancy and criminal activity (Heckman et al., 2010; Walker et al., 2011; Gertler et al., 2014; Campbell et al., 2014).

Despite this evidence that at-scale parenting programs can improve child outcomes in the short-run at their conclusion and that more intensive interventions can lead to substantial long-run impacts, an unresolved question is what mechanisms drive sustained impact. Studies that have followed children in the years following the conclusion of parenting programs find substantial variation in medium run effects. Some studies have found improved outcomes in the medium run, while others have show large increases in skills at the conclusion of interventions, followed by rapid fade-out over time (Barnett, 2011; Bailey et al., 2017). In low-and middle-income countries, two recent studies, for instance, have found different results in medium-run follow-ups. A follow-up study of a parenting intervention in Pakistan found initial improvements in early skills of two-year-old infants to persist two years after program completion (Yousafzai et al., 2016). In contrast, a follow-up of a similar intervention in Colombia found that initial gains in early skills fade-out two years after that program (Andrew et al., 2018). To reconcile these findings more evidence is needed on the factors that are driving variation in the sustained benefits of these programs. Such evidence may inform the design of interventions to produce more sustainable gains.

In this study, we attempt to shed some light on this question by exploring the effects of a parenting program in rural China two and a half years after completion as children enter schooling. Loosely modeled on the Jamaican Ready to Learn intervention (Grantham-McGregor et al., 1991), the program promoted ECD in rural China through a home-based parent training intervention implemented by officials associated with China's Family Planning Commission (FPC) (Sylvia et al., 2020). At the end of the initial intervention when the treatment and control children were 30 to 42 months old, Sylvia et al. (2020) measured substantial increases in cognitive skills in children assigned to receive weekly home visits. The study also found that the improvements in infant skill development were accompanied by increases in both parental investment and parenting skills, particularly among children who were more disadvantaged at baseline.

Two and a half years after the initial intervention, we find large persistent effects on child working memory - a key skill of executive functioning which plays a central role in children's cognitive functioning, behaviour, emotional control, and social interaction. We further find that the observed increases in parental investment behaviour at the completion of the ECD intervention persisted two and a half years later as parents in the treatment villages continue to spend more time with their children. Beyond parental investments in the home, we also find that children in treatment villages enrol earlier and in better quality preschools. Moreover, changes in caregiver schooling decisions reflect an increase in valuation of school quality relative to other attributes. Our finding of persistent effects on caregivers' preschool enrolment decisions points to an important intervention-induced persistent shift in investment behaviour which might lead to long-term benefits over the life-cycle.

Our results imply an important role for preschool quality in sustaining longer-term effects of early childhood interventions, particularly where caregivers have low levels of human capital. Previous research in diverse settings has shown that it matters when and where parents decide to enrol their children in school. The age at which children start school affects both short- and longer-term educational attainment as well as health and labour market outcomes (Elder and Lubotsky, 2009; Carlsson et al., 2015). Several studies show that the returns to preschool are substantial, especially for disadvantaged children who typically have less stimulating counter-factual home environments (Havnes and Mogstad, 2015; Felfe et al., 2015; Herbst, 2017; Cornelissen et al., 2018). Recent evidence suggests that not just the quantity but, more importantly, the quality of preschool is key to child development (Araujo et al., 2016; Andrew et al., 2019; Nores et al., 2019). Araujo et al. (2016) find that preschool quality has positive effects on student learning and executive functioning skills at school entry in Ecuador. Both Andrew et al. (2019) and Nores et al. (2019) find that improvements in preschool quality led to significant improvements in cognitive and language skills for preschool children in Colombia. Other studies find that returns to preschool quality are persistent and lead to increased earnings, higher college attendance and lower teen pregnancy (Chetty et al., 2011, 2014).

Our findings contribute to the literature on early childhood programs, particularly the "puzzle" of intervention fade-out in the medium run but evidence of long run effects on labor market outcomes and health in adulthood (Barnett, 2011; Bailey et al., 2017). Several hypotheses have been put forward in the literature to explain this apparent puzzle. It is possible that the treatment intensity typical in smaller efficacy trials is difficult to replicate in larger-scale integrated interventions (Andrew et al., 2018). It also is plausible that short-term intervention effects might appear to fade in the medium run, but then re-appear later in the life-cycle because medium-run evaluations are not measuring the relevant skills (Heckman et al., 2013) or they are measuring the relevant skills, but are doing so with large measurement error (Cunha et al., 2010; Johnston et al., 2014; Laajaj and Macours, 2017). Measurement error would make it more challenging to observe medium-term intervention effects if initial skill differences are small but nevertheless accumulate over time and end up resulting in long-term differences in health and labour market outcomes.

Our analysis highlights another potential explanation for mixed findings on medium-term effects: that ECD intervention programs differ in their ability to permanently shift investment behaviour. For the few early childhood interventions that have followed children, there has often been little information collected on subsequent investments from parents or schools. Two notable exceptions are Andrew et al. (2018) and Yousafzai et al. (2016), who collect information on parental investment behaviour two years after the end of the parenting interventions. Andrew et al. (2018) find that initial improvements in time and material investment observed at the end of the parenting intervention had disappeared. Yousafzai et al. (2016), on the other hand, found persistent improved parenting practises and more stimulating home environment two years after the intervention had ended.

In contrast to parental investments at home, however, how interventions affect subsequent decisions regarding school enrolment may be more fundamental to sustaining longer-run effects, particularly among children who would be more likely to have low levels of skills absent intervention. Some previous evidence for this comes from Currie and Thomas (2000) who find that initial gains in test scores for children who attended Head Start in the U.S. faded out more rapidly for black students compared to whites because they were more likely to subsequently select in to lower-quality schools. Distinguishing between parental home investments and investments through schooling decisions may help to explain the apparent persistence puzzle if the short-term improvements in skill formation observed in the literature are mainly driven by the increased cognitive stimulation during the intervention while the long-term health and labour market benefits result from a shift in parental investment behaviour after the intervention. As children age, schooling is likely to become relatively more important to skill development than parental investments of time and resources at home. And for parents who believe that investments are important but perceive investment of their own time to have low returns, the perceived net benefits of schooling may become larger at an an earlier age and increase with school quality.

The remainder of the paper is structured as follows. In the next section, we discuss some stylized facts and findings from the literature on preschool enrolment in China and beyond. In section 3 we describe the experimental design and data collection. In section 4 we report the findings of the medium-run follow-up impact evaluation of the parenting intervention. Section 5 concludes.

2 Stylized Facts and Findings on Preschool Enrolment in China and Beyond

Chinese parents can enrol their children in preschool between the ages of three and six years old before entering compulsory primary school. Preschool enrolment is not compulsory but parents are increasingly deciding to enrol their children in school early as evident in the rapidly increasing national preschool enrolment rates. From 2009 to 2016, the gross national preschool enrolment rate increased from 50.1% to 77.4% (Wu, 2017).[1] In 2010, the central government declared that by 2020, 95% of Chinese children should receive at least one year of preschool education and 75% of children should receive three years of preschool education (MOE, 2010).

Despite these promising improvements in national preschool enrolment rates, less is known about preschool enrolment in the rural parts of China, where 63.78% of China's populations live (NBS, 2015). More generally, increased preschool enrolment rates are not a guarantee for improvements in human capital. Research on the returns to preschool show mixed results. Several studies find the returns to expansion of universal preschool to be positive (Berlinski et al., 2008, 2009; Havnes and Mogstad, 2011; Felfe et al., 2015) whereas other studies find mixed or no impacts (Magnuson et al., 2007; Gupta and Simonsen, 2010) or even negative impacts on child development (Baker et al., 2008, 2019).[2] This mixed evidence might result from substantial heterogeneity in the quality of alternative care environments. Disadvantaged

[1] Enrolment rate is the ratio of the number of children in preschool to the number of children between 3-6 years old.

[2] Elango et al. (2015) provides an extensive review of this literature.

children typically have less stimulating home environments and might therefore benefit more from universal preschool than the average child. Several studies indeed find that returns to universal preschool are higher for disadvantaged children (Havnes and Mogstad, 2015; Felfe et al., 2015; Herbst, 2017; Cornelissen et al., 2018).

Another important source of heterogeneity in returns to preschool is the quality of the learning environment in preschools (Ulferts et al., 2019). Araujo et al. (2016) find that preschool quality has positive effects on student learning and executive functioning skills at school entry. Other studies find that returns to preschool quality are persistent and lead to increased earnings, higher college attendance and lower teen pregnancy (Chetty et al., 2011, 2014). Taken together, this evidence shows that the effectiveness of preschool depends on the counterfactual care arrangement the program is substituting for and the change in learning quality it presents (Cascio, 2015).

The importance of preschool quality might therefore be even more pressing in the context of low and middle income countries as a large share of children grow up in insufficiently stimulating home environments resulting in major risk of cognitive delays at school entry (Lu et al., 2016; Richter et al., 2017). One concern is that there might be large differences in the quality of the learning environment between rural and urban preschools in China. Although there is generally little research into rural preschools, one study in Shaanxi province found that in most rural preschools, only the principals had participated in city-level teacher training, and most teachers have never participated in any training. Moreover, preschool teachers in rural areas tend to be young and the share of experienced teachers is small (Lai et al., 2015).

Even within the rural areas in China, there is reason to expect considerable heterogeneity in preschool quality between different administrative levels.[3] Generally, local township governments tend to fund one to two public preschools within each township whereas most village preschools are privately owned. While most township preschools get the majority of their funding from the government, village preschools only receive a minimal subsidy per headcount (Wu and Qin, 2012, 2016). These different funding channels lead to substantial

[3]There are five levels of local government: the provincial, prefecture, county, township, and village. Townships and villages are the most relevant local administrative levels in rural areas.

quality differences between township and village preschools. Wu and Qin (2012) find that in general village preschools have less qualified teachers, a lower teacher-pupil ratio and lower quality preschool facilities.

As part of this study we collected detailed information on all preschools that had one or more of the sample children enrolled at the time of the survey. Enumerators were sent to each school to survey both teachers and head teachers on the structural and process quality of the preschools. These unique data allow us to study in detail the parental preschool decision in this rural part of China and we present our findings in section 4.

3 Experimental Design and Data Collection

3.1 Sampling and Randomisation

The study sample was selected from one prefecture located in a relatively poor province located in Northwest China. The province ranks in the bottom half of provinces nationally in terms of GDP per capita. The prefecture chosen for the study is located in a mountainous and relatively poor region of the province. Nearly all of the people residing in this prefecture are ethnically Han.

The research team used a systematic protocol to select the sample. As the first step researchers selected townships from four nationally-designated poverty counties in the chosen prefecture. All townships in each county were included except the one township in each county that housed the county seat. Within each township, administrative data were used to compile a list of all villages reporting a population of at least 800 people. Next, two villages were randomly selected from the list in each township. These exclusion criteria were applied to ensure the sample villages had a sufficient number of children in the target age range. All children in sample villages between 18 and 30 months of age were enrolled in the study.

Before the start of the intervention sample villages were randomly assigned to a treatment (n=65) and control arm (n=66) and the randomisation procedure was stratified at the county level. Next, each parenting trainer was assigned a maximum of four families chosen randomly from treatment villages to be enrolled in the program. In treatment villages, this resulted in a sample of 212 children enrolled in the parenting intervention and a remaining 79 that were not. In control villages, a total sample of 300 children was enrolled.

3.2 Parenting Program

The parenting intervention was a weekly home visiting program where parenting trainers trained caregivers to interact with their offspring through cognitively stimulating and developmentally-appropriate activities using a structured curriculum. The teaching curriculum is based loosely on the Jamaican home visiting model (Grantham-McGregor et al., 1991) and adapted by child development psychologists in China to the local rural setting. During the weekly sessions, parenting trainers would introduce main caregivers (typically, mother or grandmother) to the activity and assist caregivers to engage in the activity with their child. At the end of each weekly session, the materials used for that week's activities (toys and books) were left in the household to be returned at the next visit.

The 6-month long parenting intervention started in November 2014 and ended in April 2015. The parenting trainers, selected by the FPC from among their cadres in each township received an initial, one-week intensive training at the beginning of the program which covered theories and principles of early childhood development, parenting skills, and the curriculum. This initial training consisted of both classroom-based instruction as well as field practice. Throughout the program, trainers received periodic training by phone on curriculum activities which would vary according to the ages of children to whom they were assigned. For more details about the intervention, refer to Sylvia et al. (2020).

3.3 Data Collection and Measurement

A team of enumerators conducted a baseline survey in October 2014 and a follow-up survey in May 2015 at the end of the parenting intervention.[4] In August 2017, we conducted a follow-up survey of all the households enrolled in the initial randomised controlled trial (See Figure 1).

[4]Details on the baseline and endline data collection procedure can be found in Sylvia et al. (2020)

Enumerators collected detailed information on infant skills, parental investment behaviour, and a wide range of household and preschool characteristics.

3.3.1 Measuring Infant Skills

Infant skills are assessed using a battery of questions from the Weschler Preschool and Primary Scale of Intelligence (WPPSI) and the Strengths and Difficulties Questionnaire (SDQ). The Weschler Preschool and Primary Scale of Intelligence (Wechsler, 2012) is designed to measure cognitive development of pre-schoolers and young children. For this study we use the WPPSI-IV edition that has been translated into Chinese (Wechsler, 2014) and administered to children by trained enumerators. WPPSI-IV measures five main skill domains:

i *Verbal Comprehension Index:* measures verbal comprehension and reasoning skills.

ii *Visual Spatial Index:* measures the ability to organise and understand visual parts and information, assimilate visual and motor functions simultaneously, and see the whole-part connection to objects.

iii *Fluid Reasoning Index:* measures the ability to utilize inductive reasoning (e.g. use past observations to predict current situations)

iv *Working Memory Index:* measures the ability to balance focus and attention while manipulating visual and auditory information in conscious awareness.

v *Processing Speed Index:* measures how fast a child can scan and differentiate visual information.

Conceptually, WPPSI-IV is developed as an IQ test but in practice several sub-domains measure executive functioning (EF) skills as well. The Processing Speed Index of WPPSI is related to the EF sub-domain of Information Processing whereas the Working Memory Index of WPPSI-IV is related to the domain of Cognitive Flexibility of EF. We are not aware of any study specifically linking WPPSI-IV to early executive functioning skills and therefore assume that we are measuring a subset of executive functioning domains as well as cognitive development more generally.

To measure domains (i)-(iv), all administered tests items increase in difficulty and the test is stopped when the child can no longer provide a correct answer. Given this specific test structure, a simple average score of all correctly answered items would provide a noisy measure of underlying child ability. If all items are of identical difficulty than the simple average is the best estimate of a child's underlying ability. However, to the extent that items differ in their difficulty level, then a weighted average can provide a more precise estimate of child ability by assigning higher weights to more difficult items. Hence, in a first step, we estimate a two-parameter logistic IRT-model which calculates the optimal weighted average of all items taking into account response patterns. To fix ideas, we briefly discuss how an IRT measurement system can help mitigate measurement error:

Let I_{ij}^{λ} define the performance measure for child i and item j on test λ and let's assume it is determined as follows:

$$I_{ij}^{\lambda} = \beta_j + \alpha_j \Lambda_i^{\lambda} + \epsilon_{ij}^{\lambda} \tag{1}$$

where Λ_i^{λ} is child i's latent skill for test λ and this is assumed to be independent from the error term ϵ_{ij}^{λ}. In other words, we assume that a unidimensional skill is sufficient in explaining a child's response behaviour on items in each sub-test We further assume that a child's response to an item is independent of his or her responses to other items after conditioning on child latent skill.

The variable I_{ij}^{λ} is not observed to the enumerator or caregiver. Instead, we observe I_{ij}^{λ}=1 if $I_{ij}^{\lambda} > 0$ and $I_{ij}^{\lambda} = 0$ otherwise. We further assume that the measurement system is invariant to treatment assignment. The estimated item specific intercepts, $\hat{\beta}_j$, represent the level of difficulty of item j. The estimated parameter $\hat{\alpha}_j$ represents the discrimination ability of item j. Hence, in the 2-parameter logistic IRT model, the probability of success on an item j is a function of both the level of latent skill Λ_i^{λ} and the difficulty level and discrimination ability of item j. We refer to Appendix B1 for a more details on the two-parameter logistic IRT measurement model.

Each of the first four indexes (i)-(iv) of WPPSI-IV are administered by two separate sub-tests which are described in more detail in Appendix B. We use the raw item data for each sub-test λ to estimate a 2-parameter logistic IRT model as specified above.[5] We estimate parameters α_j and β_j using maximum likelihood and integrate out the latent skill Λ_i^λ. In a second step we use empirical Bayes estimators of the latent skill Λ_i^λ and take the mean of the empirical posterior distribution of Λ_i^λ , conditional on child's item responses I_j^λ, while imposing the estimated parameters $\hat{\alpha_j}$, $\hat{\beta_j}$. Figures A1- A4 in Appendix A plot kernel density estimates of latent skill Λ_i^λ for each sub-test λ in WPPSI-IV index (i)-(iv). Next, we rank performance on each sub-test λ based on the estimated skill factor scores $\hat{\Lambda_i^\lambda}$. The final score for each of the four WPPSI-IV indexes (i)-(iv) is calculated as the average rank-performance on the two sub-tests.

The last index of WPPSI-IV, (v) Processing Speed, is administered by two tests that calculate the amount of time it takes a child to complete a task and therefore is not dependent on increasing task difficulty as is the case for the four other WPPSI-IV indexes. For this index, the score is calculated as the average time of test completion on both sub-tests. All skill factors for index (i)-(v) are standardised non-parametrically for each age-month group as infant skills mature rapidly over time.[6] Kernel density estimates of the five infant skill distributions for control and treatment villages are plotted in Figure 2.

The Strengths and Difficulty Questionnaire (Goodman et al., 2000) is a 25-item carer-reported instrument for the assessment of social, emotional, and behavioural functioning of children and adolescents ages 2 to 17 years old. For this study, we use an SDQ questionnaire that was translated and validated for the Chinese context (Du et al., 2008). SDQ measures the following three non-cognitive sub-domains:

i *Externalising Behaviour:* measures behavioural problems that are manifested in children's outward behaviour such as disruptiveness, hyperactivity, and aggressive behaviour.

[5]Items with zero variance are excluded from the analysis and represent less than 1% of all item measures.

[6]This standardisation method is less sensitive to outliers and small sample size within age-category and gives us normally distributed internally standardised scores with mean zero across each age-month group (Attanasio et al., 2015).

ii *Internalising Behaviour:* measures behavioural problems affecting children's internal psychological environment such as withdrawn, anxious, and depressed behaviour.

iii *Pro-social Behaviour:* measures positive behaviours, attitudes, and emotions directed towards others.

All items for indexes (i)-(iii) are scaled on a 3-point likert scale (1 *not true*, 2 *somewhat true*, 3 *certainly true*). We estimate a dedicated measurement system relating all observed items to a latent factor capturing the above three SDQ sub-domains:

$$I_{ij}^{\lambda} = \mu_j + \gamma_j \theta_i^{\lambda} + \delta_{ij}^{\lambda} \tag{2}$$

with I_{ij}^{λ} the observed j^{th} measure for child i; μ_j the mean of the j$^{\text{th}}$ measure and γ_j the loading of the factor for measure j. The measurement error δ_{ij}^{λ} is the remaining proportion of the variance in measure j that is not explained by the latent non-cognitive skill factor θ_i^{λ} and assumed to be independent and have a zero mean. We further assume again that the measurement system is invariant to treatment assignment. The parameters of the measurement system are estimated using maximum likelihood and can be found in Table B1 in Appendix B.

We next use the estimated means and factor loadings from (3) to predict the three latent non-cognitive skill factors, θ_i^{λ} , for each child i in the sample using the Bartlett scoring method (Bartlett, 1937). The predicted non-cognitive skill factors are standardized non-parametrically for each age-month group. Kernel density estimates of the three non-cognitive skill distributions for control and treatment villages are plotted in Figure 3.

3.3.2 Measuring Parental Investment

Parental investment is measured on several dimensions. First, we measure the parent's decision to enrol his/her child into preschool, at what age parents decide to enrol their child, and the quality of the preschool selection. Enrolment rates and the enrolment age distribution for control and treatment villages are plotted in Figure 4 in panel (a) and panel (c) respectively. The preschool enrolment decision by rural Chinese parents might present an important investment channel. We collect information on 165 preschools in which children from control and treatment villages were enrolled at the time of the survey. From the 429 children that enrolled in preschool, 221 are enrolled in village preschools and 208 in township or county preschools.[7]

Enumerators were sent out to survey both teachers and headteachers to collect information on preschool characteristics, teacher characteristics as well as measures of structural and process quality. Preschool characteristics are measured by the number of enrolled pupils, the share of pupils receiving government need-based aid, and the tuition fees paid by parents each semester.[8] Teacher characteristics are measured by age, gender, years of experience on the job, salary, and educational attainment as well as whether teachers have received any professional training in the past year.

Structural quality of preschools is measured by the pupil-teacher ratio, the number of activity rooms, the size of the outdoor play area, and whether the preschool has a designated playroom, exercise room, dormitories and provides breakfast to pupils. To assess process quality we collect information on whether teachers engage in a set of teaching activities such as: reading books in class, organising physical exercise activities, art&music activities, and science activities. We also collect information on whether social, and language skills are taught in class.

We use exploratory factor analysis (EFA) to guide us in the dimensionality reduction of the preschool quality data and determine the optimal factor structure (Appendix B.4.). We derive a one-factor model which captures general preschool quality. The pattern of estimated factor loadings (Table B5) shows that higher preschool quality factor scores are associated with with schools that are bigger in size, more likely to be located in township or counties as compared to villages and have younger and more educated teachers that are more likely to have received a teacher training in the past year.[9] Higher factor scores are also associated with larger indoor and outdoor space, the availability of dormitories and breakfast and several measures of process

[7]We were not able to obtain the preschool information for 22 households who enrolled at preschools that are out of enumeration area.

[8]In 2016, the need-based aid in survey region covers up to 750 Yuan per year, equalling to about 110 USD

[9]A county in China usually consists of several townships and each township usually resides more than 10 villages. Township or county preschools typically offer higher process and structural quality hence the decision to enrol offspring outside the village in a county or township preschools potentially captures an important parental investment channel in rural China (Zhao and Hu, 2008; Hu et al., 2014)

quality such as organising exercise and science activities and reading books in class.

Secondly, we measure parental investment at school-entry by asking parents how much time and money they spend on their children. The main caregiver is asked whether he/she engaged in a set of child-rearing activities the previous day, including story-telling, singing songs, interactive play activities, and how long on average the child spends watching TV during the day. Caregivers are also asked to report how much money they spend on children's books, toys, clothes and school expenditures in the last year.[10] Exploratory factor analysis indicates both time and material parental investment are best measured using a one-factor measurement system (Appendix B.3.). Estimated factor loadings of the measurement system can be found in Table B5. Both investment factors are standardized by the distribution of the control group.

3.4 Summary Statistics, Balance, and Attrition

Summary statistics and tests for balance across control and treatment groups during baseline are shown in Table 1. Differences between study arms in individual child and caregiver characteristics are insignificant. A joint significance test across all baseline characteristics also confirms that the study arms are balanced.[11]

Children in our sample are on average just over 24 months old at the start of the program. Less than 5% of the sample children are born with low birth weight. A large part of the children in our sample are firstborn in the family (60%). More than 80% of children were ever breastfed and around 35% were breastfed for more than one year. More than 20% percent of sample children were anemic according to the WHO-defined threshold of 110 g/L. On average children were reported to be ill 4 days over the previous month.[12] At baseline, around 40 percent of the sample is cognitively delayed with Bayley MDI scores below 80 points, but few (10%) were delayed in their motor development. Around 30 percent of the children are at risk of social-emotional problems at baseline.

[10] School expenditures include both annual tuition fees and other school-related expenditures.

[11] We test this by regressing treatment status on all baseline characteristics reported in Table 1 and test that the coefficients on all characteristics were jointly zero. The p-value of this test is 0.529.

[12] Caregivers were asked whether the child had suffered from fever, cough, diarrhoea, indigestion, or respiratory cold over the previous month.

We also collected information on caregivers and families. Around 26 percent of the sample receives social security support through the *dibao*, China's minimum living standard guarantee program, as reported in Panel B of Table 1. The biological mother is the primary caregiver in only 60 percent of households, with grandmothers often taking over child-rearing when mothers out-migrate to join the labor force in larger cities. We find that slightly more than 70 percent of primary caregivers in the sample (mothers or grandmothers) have at least 9 years of formal schooling. On average households reported being somewhat indifferent in their feelings toward the Family Planning Commission at baseline.

Baseline statistics on parental inputs, shown in Panel C of Table 1, demonstrate that caregivers engage in few stimulating activities with their children. Only 11% of caregivers told a story to their child the previous day. Less than 5% read a book to their child (on average households have only 1.6 books). Only around 1 in 3 caregivers report playing with or singing to their child the previous day.

Overall attrition between November 2014 and May 2015 was less than 1 percent and insignificantly correlated with treatment status. We define attrition as missing a Bayley's or Griffith outcome (depending on the age-cohort) measure at endline for children with a Bayley baseline measure. Attrition at follow-up two and a half years after program completion was 7.4% and balanced between treatment and control group. Table A1 in the Appendix A shows summary statistics of baseline characteristics of the non-attrited sample and shows that our follow-up sample is also balanced on baseline characteristics.

3.5 Estimation of Program Effects

We estimate the medium-run impact of the 6-month parenting intervention on our measures of infant skills and parental investment. Given the random assignment of households into treatment and control groups, comparison of outcome variable means across treatment arms provides unbiased estimates of the effect of the parenting intervention on outcomes. However, to increase power (and to account for our stratified randomisation procedure) we condition our estimates on randomisation strata (Bruhn and McKenzie, 2009) and baseline values of the outcome variable.

We use ordinary least-squares (OLS) to estimate the intention-to-treat (ITT) effects of the parenting intervention with the following ANCOVA specification:

$$Y_{ijt} = \alpha_1 + \beta_1 T_{jt} + \gamma_1 Y_{ij(t-1)} + \tau_s + \epsilon_{ij} \tag{3}$$

where Y_{ijt} is an outcome measure of child i or the parental investment from household i in village j at follow-up; T_{jt} is a dummy variable indicating the treatment assignment of village j; $Y_{ij(t-1)}$ is the outcome measure for child i at baseline, and τ_s is a set of strata fixed effects. We adjust standard errors for clustering at the village level using the Liang-Zeger estimator.

3.6 Mediation Analysis of Medium-Run Treatment Impact

An important policy question in the ECD literature is how benefits of parenting interventions are sustained over the life-cycle. We therefore conduct a mediation analysis and test whether the intervention treatment impact observed in the medium run is driven by the direct program effect of the 6-month ECD intervention and/or mediated by a shift in parental investment behaviour as a result of the intervention. The rich data we collected on preschools quality allows us to test for mediation effects beyond parental time and material investments at home and also consider the parental preschool enrolment decision.

We perform a mediation analysis following Heckman et al. (2013) in which we add the mediation channels of interest, the three parental investment factors, to equation (3) and estimate the following specification:

$$Y_{ijt} = \alpha_1 + \beta_1 T_{jt} + \eta I_{ijt} + \gamma_1 Y_{ij(t-1)} + X_{ijt}\gamma_2 + \tau_s + \epsilon_{ij} \tag{4}$$

where I_{ijt} measures parental investment behaviours at the medium-term follow-up survey. In a first specification we include each parental investment channel separately. Next, we control for all mediation channels simultaneously. In our all specifications we include a wide range of covariates X_{ijt} to account for shocks that potentially both affect the skill formation of children and our mediating investment factors.

In order to attribute a causal interpretation to the mediation analysis we need to assume that the mediating investment factors are exogenous with respect to infant skills conditional on other observed covariates and treatment assignment . However, it is possible that unobservable shocks affect both the skill formation of children and mediating investment factors, which would result in biased estimates. For example, negative shocks to the health of children could both delay the skill formation and preschool enrolment of children. We address this concern in two ways. First, we identify mediating factors over a theoretically comprehensive set of parental human capital investments by including measures of parental investments in both the home and school environment which reduces the possibility of confounding variables that would introduce bias in the estimation (Heckman and Pinto, 2015).

Second, we implement unobservable selection and coefficient stability tests proposed by Oster (2019) to test if coefficients on mediating investment factors change based on the inclusion of observable controls.[13] If the coefficient of the mediating factors does not change with the inclusion of observable controls, Oster (2019) argues that coefficient estimates are unlikely to change with the inclusion of unobserbable variables. Hence, we calculate bounds for the coefficients of the mediating investment factors for two potential scenario. In a first scenario we assume that there is no unobservable selection bias ($\delta = 0$), whereas in the second scenario we assume the bias resulting from unobervables is equal to the bias from observable controls ($\delta = 1$). Next, we report a statistic known as Oster's δ for each estimated coefficient of our mediating investment factors, which represents the ratio of the selection bias introduced by unobservables compared to the bias introduced by the observable controls that is necessary for the indirect treatment effects of the mediating factors to be statistically insignificant.[14] Hence, for high

[13] Bloem and Wydick (2019) implements the same procedure to test the robustness of a mediation analysis of a kindergarten intervention in the Philippines.

[14] We assume $R^2_{max} = 1.3\ \tilde{R^2}$, as suggested by Oster (2019). To acquire $\tilde{R^2}$, we implement a LASSO regression with cross-validation to find the optimal model fit for the WPPSI working memory index. We find that $\tilde{R^2}$ equals 0.31 and hence set R^2_{max} equal to 0.403.

values of Oster's δ we can be more confident that the estimated mediation treatment effects are robust to selection bias.[15]

4 Medium-Term Impact of the Parenting Intervention at School-Entry

4.1 Average Treatment Effects on Infant Skills

Medium-run average treatment effects of the parenting intervention measured on cognitive and non-cognitive skills at school entry can be found in Tables 2 & 3. We find that the 6-month parenting intervention led to a 0.27 standard deviation increase in working memory two and half years after program completion. For the other sub-domains of the WPPSI and measured non-cognitive skills, we find no significant differences between the control and treatment villages.

Kernel density graphs of all latent skill measures by treatment status are shown in Figure 2. The working memory skill distribution is clearly shifted to the right for the treatment group and shows improvements have been achieved across the entire ability distribution. A Kolmogorov-Smirnov (K-S) test rejects the equality of working memory skill distribution in the treatment and the control group with a p-value of 0.004.

4.2 Average Treatment Effects on Parental Investment

We find that the parenting intervention had lasting effects on parental investment. At the time of the endline survey, children in the treatment group continued to receive higher levels of investment at home. Additionally children in the treatment group were significantly more likely than the control group to be enrolled in preschool, be enrolled in preschool at earlier ages, and to be enrolled in higher-quality preschools.

In Panel A of Table 4 we report the average treatment effects on parental time and material

[15] A negative Oster's δ indicates that the observables are positively correlated with the mediators and that the unobservables have to be negatively correlated with the mediators to get null results

investment at school-entry. At the end of the parenting intervention, we had observed large increases in parental time investment as a result of the intervention (0.69 SD) Sylvia et al. (2020). We find evidence that this effect persisted, with the parental time investment factor remaining 0.30 standard deviations higher for the treatment group. The estimated medium-run effect on material investments is positive (0.10 standard deviations), but does not reach conventional levels of significance.

Panel B of the same table reports effects on preschool enrolment. We find that parents in treatment villages were approximately 7 percentage points more likely to have enrolled their children in preschool by the time of the survey. At the time of the endline survey, around 10% of children in our sample were not (yet) enrolled in preschool but rates of enrolment were significantly higher in treatment villages.In treatment villages around 5% were not (yet) enrolled in preschool as opposed to 13% in control villages (Figure 4a).

We further find that parents in treatment villages also enrolled their children at younger ages. Children in treatment villages were on average enrolled 2 months earlier than children in control villages.[16] Finally, the intervention led parents in treatment villages to enrol their children in higher quality preschools. We find that children from treatment villages are enrolled in preschools scoring on average 0.28 SD higher in terms of the preschool quality index. Figure 4d plots the preschool quality distribution for control and treatment villages. The distribution of the preschool quality factor for treatment villages is stochastically dominant and a Kolmogorov-Smirnov (K-S) test rejects the equality of the two factor distributions with a p-value < 0.001. [17]

[16] As around 10% of children were not yet enrolled in school at the time of the survey, we have missing information on their age of future enrolment. Therefore, we performed a bounding exercise to estimate the program treatment effects on the age of preschool enrolment. We first assume that all children are enrolled latest by the maximum age in the sample distribution which gives us an upper bound estimate of the treatment effect as presented in Table 4 (point estimate: -1.970, std error: 1.016). Next, we assume all remaining children enrol directly after the survey takes place, which allows us to estimate a lower bound treatment effect (point estimate: -1.573, std error: 0.911). Both findings are statistically significant at conventional levels.

[17] Similarly, we bound the estimates of the program treatment effects on preschool quality. We first assume that all remaining children are enrolled in worse quality preschool in the township where the households reside which gives us an upper bound estimate of the treatment effect (point estimate:0.357, std error: 0.132). We assume all remaining children are enrolled in best quality preschool in the township where the households reside which gives us an lower bound estimate of the treatment effect (point estimate:0.204, std error: 0.126). The former is significant at the 1% level and the latter is marginally significant at 10% level.

Taken together, these results indicate important shifts in the levels of parental investment behaviour lasting two and a half years after completion of the intervention. In addition to higher (parent-reported) investments at home, there were meaningful shifts in the parental preschool enrolment decision attributable to the intervention.

4.3 Parental Preschool Selection Decision

The investments parents make in the school environment of their children may be fundamental in sustaining longer-run impacts parenting interventions, particularly among children with high prevalence of cognitive delay. We therefore conducted additional analyses to further explore how the parenting intervention impacted parental schooling decisions. Our survey design allows us to study the parental preschool decision in detail as we have information on both the location of the households and the preschools that children enrol in.[18] In this section, we present estimates of the effect of the intervention on school choice and explore how the intervention shifted parental schooling preferences, specifically how parents weight alternative school attributes.

4.3.1 Village vs. Township Preschool Enrolment

At the time of the survey, 90.5% of sample children were enrolled in a preschool. Among the children enrolled, 55.9% of them attended a local village preschool while 44.1% attended preschools in townships or county seats. Figure 5 shows the geographical location of the preschools in the study region and maps the villages that were enrolled in the original randomized experiment. The figure also shows the enrolment rates in township or county preschool and village preschools for treatment and control villages.

There are no administrative obstacles to enrol children in preschools within the county of residence, but enrolment across county lines can be costly both due to longer travel times and restrictions due to the the Hukou system (Randau and Medinskaya, 2015). In our sample we find that only 0.4% of children enrolled in preschools outside their county of residence so for

[18]Location data on preschools is scraped through Gaode Map API.

the rest of this analysis we abstract from this and focus on within county preschool choices.

We begin our analysis by formally estimating program treatment effects on the decision to enrol children in county or township preschools versus village preschools (Figure 4b and Table 5). Preschools located in villages differ from preschools in townships along several important dimensions. Among other differences, they tend to be larger, have fewer students receiving need-based government aid, have teachers with better formal qualifications, pay their teachers higher salaries, have superior facilities, and report practices in line with better process quality (see Appendix Tables A2a-A2b). They also tend to charge around two times the amount of tuition: Village preschools charge on average 608 yuan (\$90 at the time of the survey) per semester compared to 1276 yuan (\$189) per semester in township preschools.

Despite significantly higher costs of attendance, we find that around 60% of children in the treatment villages enrolled in township preschools compared to 40% in the control villages (Figure 4b). We estimate a treatment effect on enrolment in township preschools of 18.4 percentage points. In line with this, we also estimate that households in the treatment group were 18.2 percentage points less likely to enrol their children in the preschool that was closest to them (Table 5). [19]

4.3.2 Schooling Preferences

The treatment effects that we find on enrolment in preschool and likelihood of enrolling in township over village preschools suggest that the intervention affected not only the decision of whether to attend preschools, but also shifted how parents value different attributes of schools in making schooling decisions. We explore this further by estimating households' preferences over three preschool attributes —total preschool expenses (tuition + fees), preschool quality, and distance to preschool —and how these differ between the treatment and control groups.

Similar to Burgess et al. (2015) and following McFadden's approach, we estimate a condi-

[19] See Appendix C for additional information on preschool choices and preschool enrolment of the sample.

tional logit model, assuming that error terms have standard Type 1 extreme value distribution:

$$Enrollment_{is} = \frac{e^{X_{is}\beta + T_i X_{is}\beta_t}}{\sum_{l=1}^{n} e^{X_{il}\beta + T_i X_{il}\beta_t}} \quad s = 1, ...n$$

where X_{is} is the preschool attributes, T_i is the treatment assignment of individual i. To construct preschool choice sets, we match households with all preschools in their own county and drop school choices that are more than 81 kilometres from households (we don't observe any preschool enrolment above that distance.) For simplicity, we also drop the 10% of the sample children who were not enrolled in any preschool at the time of the survey.

The results of this analysis are reported in Table 6. The first two columns in the table show estimates from the conditional logit model and the remaining four columns show the implied demand elasticities. We find that school choice is inelastic with respect to tuition and fees for both treatment and control households and that choices are similarly highly elastic with respect to distance. Treatment households, however, appear to place a significantly higher value on preschool quality. The coefficient on the interaction term between treatment and school quality in the conditional logit model indicates a large and significant difference in the weight given to preschool quality in the treatment and control groups. We estimate an elasticity with respect to preschool quality of 1.33 in the treatment group compared to 0.17 in control villages.

Taken together, these results provide some evidence that the parenting intervention not only increased preschool attendance, but also led to a shift in schooling preference two and a half years after the conclusion of the intervention.

4.4 Mediating Medium Run Program Impacts: Parental Investment at Home vs. Preschool

Two and half years after the initial parenting intervention we find large persistent program impacts on infant working memory skills and parental investment behaviour. A remaining question is to what degree this persistence in effects on infant skill accumulation was mediated through a shift in parental investment behaviour relative to other channels. In Table 7 we present our findings of the mediation analysis as described in section 3.6. In column (1) we first report the average program treatment impact on infant working memory, controlling for a large set of baseline child and village characteristics. Next, in columns (2)-(4) we report estimated coefficients of the direct and mediated medium-run program impact for each of the parental investment factors separately. We view this analysis as exploratory as we lack a means of identifying causal mediation effects.

The results of these regressions indicate a large direct effect on infant skill accumulation due to assignment of households to the parenting program that remains largely unchanged even after controlling for the parental time and material investment factors measured at endline. The findings in column (4), on the other hand, provide some evidence that program effects in the medium run are at least partly mediated by improved parental investment in preschool quality. After controlling for treatment assignment to the parenting program we find that the preschool investment factor significantly predicts infant working memory skill. Moreover, including the preschool investment factor reduces the estimated direct impact of the parenting program from 0.25 to around 0.21 of a standard deviation of the working memory infant skill distribution. This finding is robust to simultaneously controlling for all parental investment factors as reported in column (5).

Given the lack of a credible means of estimating causal mediation effects, we examine the robustness of these findings by conducting unobservable selection and coefficient stability tests following Oster (2019). The estimated coefficients of the mediated program treatment impacts through the parental investment factors in home environment quality are not robust to unobserved heterogeneity as the estimated Oster's δs are respectively -0.11 and 0.62.[20] The Oster's δ estimated in column (4), however, implies that selection through unobservables would need to be about 1.6 times more influential than the relevant observables to render the mediating impact of the parental investment in preschool quality insignificant. In other words, the estimated mediation effect of parental investment in preschool quality is relatively robust to unobserved heterogeneity.

[20] The relevant cut-off value for robustness in the literature is 1 (Oster, 2019; Altonji et al., 2005)

5 Conclusion

This paper presents the medium-run impacts of a home-based parenting program delivered by cadres of China's Family Planning Commission on child development and parental investment at school-entry, two and a half years after program completion. We find large persistent intervention effects on child working memory - a key skill of executive functioning which plays a central role in children's cognitive functioning, behaviour, emotional control, and social interaction.[21] We further find that the observed change in parental investment at the completion of the parenting intervention persisted at the time children enrolled in preschool as parents in treatment villages continue to invest more time in their children. Beyond parental investments in the home, we also find that children in treatment villages enrol earlier and in better quality preschools —induced by a shift in schooling preferences —as well as evidence that this played a mediating role in effects on working memory. Taken together, our results imply an important role for preschool quality in sustaining longer-term effects of early childhood interventions, particularly where caregivers have low levels of human capital.

Our study has several limitations. First, the study took place in one disadvantaged rural area in northwest China and, therefore, the medium-run effectiveness of early childhood interventions in other regions or for other populations might differ. Second, we measured a wide range of cognitive and non-cognitive skills but not all tests administered might be developmentally-relevant for the population under study and may suffer from measurement error, especially a concern for caregiver-reported infant non-cognitive skill assessments. Similarly, we measure a large set of school characteristics associated with better quality in previous studies, but these may not accurately or completely reflect the specific attributes that predict whether schools produce learning gains in this context. Finally, we estimate effects two and a half years after the conclusion of the intervention and a longer-run follow-up of the children in the study will be necessary to observe whether benefits can be sustained over the life-cycle.

[21] In the medium term, the program also led to lasting effects on households' perception of the Family Planning Commission that were found in the short-run (Sylvia et al., 2020). In the medium term, households from the treatment villages were more likely to self-report trust in village cadres from the Family Planning Commission than those from the control villages—point estimate: 0.052, std error: 0.030. p-value:0.082.

References

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy*, 113(1):151–184.

Andrew, A., Attanasio, O., Bernal, R., Sosa, L. C., Krutikova, S., and Rubio-Codina, M. (2019). Preschool quality and child development. Technical report, National Bureau of Economic Research.

Andrew, A., Attanasio, O., Fitzsimons, E., Grantham-McGregor, S., Meghir, C., and Rubio-Codina, M. (2018). Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: A follow-up of a cluster randomised controlled trial in colombia. *PLoS medicine*, 15(4):e1002556.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., and Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, 131(3):1415–1453.

Attanasio, O., Meghir, C., and Nix, E. (2015). Human capital development and parental investment in india. Discussion paper, National Bureau of Economic Research.

Attanasio, O. P., Fernández, C., Fitzsimons, E. O., Grantham-McGregor, S. M., Meghir, C., and Rubio-Codina, M. (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:g5785.

Bailey, D., Duncan, G. J., Odgers, C. L., and Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of research on educational effectiveness*, 10(1):7–39.

Baker, M., Gruber, J., and Milligan, K. (2008). Universal child care, maternal labor supply, and family well-being. *Journal of political Economy*, 116(4):709–745.

Baker, M., Gruber, J., and Milligan, K. (2019). The long-run impacts of a universal child care program. *American Economic Journal: Economic Policy*, 11(3):1–26.

Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science*, 333(6045):975–978.

Bartlett, M. S. (1937). The statistical conception of mental factors. *British journal of Psychology*, 28(1):97–104.

Berlinski, S., Galiani, S., and Gertler, P. (2009). The effect of pre-primary education on primary school performance. *Journal of public Economics*, 93(1-2):219–234.

Berlinski, S., Galiani, S., and Manacorda, M. (2008). Giving children a better start: Preschool attendance and school-age profiles. *Journal of public Economics*, 92(5-6):1416–1440.

Black, M. M., Walker, S. P., Fernald, L. C., Andersen, C. T., DiGirolamo, A. M., Lu, C., McCoy, D. C., Fink, G., Shawar, Y. R., Shiffman, J., et al. (2017). Early childhood development coming of age: science through the life course. *The Lancet*, 389(10064):77–90.

Bloem, J. R. and Wydick, B. (2019). All i really need to know i learned in kindergarten? evidence from the philippines.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, 1(4):200–232.

Burgess, S., Greaves, E., Vignoles, A., and Wilson, D. (2015). What parents want: School preferences and school choice. *The Economic Journal*, 125(587):1262–1289.

Campbell, F., Conti, G., Heckman, J. J., Moon, S. H., Pinto, R., Pungello, E., and Pan, Y. (2014). Early childhood investments substantially boost adult health. *Science*, 343(6178):1478–1485.

Carlsson, M., Dahl, G. B., Öckert, B., and Rooth, D.-O. (2015). The effect of schooling on cognitive skills. *Review of Economics and Statistics*, 97(3):533–547.

Cascio, E. U. (2015). The promises and pitfalls of universal early education. *IZA World of Labor*.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9):2633–79.

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2018). Who benefits from universal child care? estimating marginal returns to early child care attendance. *Journal of Political Economy*, 126(6):2356–2409.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931.

Currie, J. and Thomas, D. (2000). School quality and the longer-term effects of head start. *Journal of Human Resources*, 35(4):755–774.

Dickey, W. C. and Blumberg, S. J. (2004). Revisiting the factor structure of the strengths and difficulties questionnaire: United states, 2001. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43(9):1159–1167.

Du, Y., Kou, J., and Coghill, D. (2008). The validity, reliability and normative scores of the parent, teacher and self report versions of the strengths and difficulties questionnaire in china. *Child and adolescent psychiatry and mental health*, 2(1):8.

Elango, S., García, J. L., Heckman, J. J., and Hojman, A. (2015). Early childhood education. In *Economics of Means-Tested Transfer Programs in the United States, Volume 2*, pages 235–297. University of Chicago Press.

Elder, T. E. and Lubotsky, D. H. (2009). Kindergarten entrance age and children's achievement impacts of state policies, family background, and peers. *Journal of human Resources*, 44(3):641–683.

Felfe, C., Nollenberger, N., and Rodríguez-Planas, N. (2015). Can't buy mommy's love? universal childcare and children's long-term cognitive development. *Journal of population economics*, 28(2):393–422.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., Chang, S. M., and Grantham-McGregor, S. (2014). Labor market returns to an early childhood stimulation intervention in jamaica. *Science*, 344(6187):998–1001.

Goodman, A., Lamping, D. L., and Ploubidis, G. B. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the strengths and difficulties questionnaire (sdq): data from british parents, teachers and children. *Journal of abnormal child psychology*, 38(8):1179–1191.

Goodman, R. (1997). The strengths and difficulties questionnaire: a research note. *Journal of child psychology and psychiatry*, 38(5):581–586.

Goodman, R., Ford, T., Simmons, H., Gatward, R., and Meltzer, H. (2000). Using the strengths and difficulties questionnaire (sdq) to screen for child psychiatric disorders in a community sample. *The British journal of psychiatry*, 177(6):534–539.

Gorsuch, R. L. (2003). Factor analysis. weiner, i. *Handbook of Psychology: Research Methods in Psychology*, pages 143–164.

Grantham-McGregor, S. M., Powell, C. A., Walker, S. P., and Himes, J. H. (1991). Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the jamaican study. *The Lancet*, 338(8758):1–5.

Gupta, N. D. and Simonsen, M. (2010). Non-cognitive child outcomes and universal high quality child care. *Journal of Public Economics*, 94(1-2):30–43.

Havnes, T. and Mogstad, M. (2011). No child left behind: Subsidized child care and children's long-run outcomes. *American Economic Journal: Economic Policy*, 3(2):97–129.

Havnes, T. and Mogstad, M. (2015). Is universal child care leveling the playing field? *Journal of public economics*, 127:100–114.

Heckman, J., Pinto, R., and Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103(6):2052–86.

Heckman, J. J., Liu, B., Lu, M., Zhou, J., et al. (2020). Treatment effects and the measurement of skills in a prototypical home visiting program. Technical report.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., and Yavitz, A. (2010). The rate of return to the highscope perry preschool program. *Journal of public Economics*, 94(1):114–128.

Heckman, J. J. and Pinto, R. (2015). Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometric reviews*, 34(1-2):6–31.

Herbst, C. M. (2017). Universal child care, maternal employment, and children's long-run outcomes: Evidence from the us lanham act of 1940. *Journal of Labor Economics*, 35(2):519–564.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

Hu, B. Y., Zhou, Y., Li, K., and Killingsworth Roberts, S. (2014). Examining program quality disparities between urban and rural kindergartens in china: Evidence from zhejiang. *Journal of Research in Childhood Education*, 28(4):461–483.

Johnston, D., Propper, C., Pudney, S., and Shields, M. (2014). Child mental health and educational attainment: multiple observers and the measurement error problem. *Journal of Applied Econometrics*, 29(6):880–900.

Knudsen, E. I., Heckman, J. J., Cameron, J. L., and Shonkoff, J. P. (2006). Economic, neurobiological, and behavioral perspectives on building america's future workforce. *Proceedings of the National Academy of Sciences*, 103(27):10155–10162.

Laajaj, R. and Macours, K. (2017). *Measuring skills in developing countries*. The World Bank.

Lai, Y., Xue, X., and Yang, R. (2015). On the allocation problems of teacher resources of pre-school education in rural areas and its optimized path. 36(03):103–111.

Lu, C., Black, M. M., and Richter, L. M. (2016). Risk of poor development in young children in low-income and middle-income countries: an estimation and analysis at the global, regional, and country level. *The Lancet Global Health*, 4(12):e916–e922.

Magnuson, K. A., Ruhm, C., and Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education review*, 26(1):33–51.

MOE (2010). Outline of the national medium and long-term education reform and development plan (2010-2020).

NBS (2015). China statistical yearbook 2015.

Nores, M., Bernal, R., and Barnett, W. S. (2019). Center-based care for infants and toddlers: The aeiotu randomized trial. *Economics of Education Review*, 72:30–43.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Randau, H. R. and Medinskaya, O. (2015). The education system: The big picture. In *China Business 2.0*, pages 179–184. Springer.

Richter, L. M., Daelmans, B., Lombardi, J., Heymann, J., Boo, F. L., Behrman, J. R., Lu, C., Lucas, J. E., Perez-Escamilla, R., Dua, T., et al. (2017). Investing in the foundation of sustainable development: pathways to scale up for early childhood development. *The Lancet*, 389(10064):103–118.

Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3):659–687.

Sylvia, S., Warrinnier, N., Luo, R., Yue, A., Attanasio, O., Medina, A., and Rozelle, S. (2020). From quantity to quality: Delivering a home-based parenting intervention through china's family planning cadres. *The Economic Journal*.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. American Psychological Association.

Ulferts, H., Wolf, K. M., and Anders, Y. (2019). Impact of process quality in early childhood education and care on academic outcomes: Longitudinal meta-analysis. *Child development*, 90(5):1474–1489.

Walker, S. P., Chang, S. M., Vera-Hernández, M., and Grantham-McGregor, S. (2011). Early childhood stimulation benefits adult competence and reduces violent behavior. *Pediatrics*, 127(5):849–857.

Wechsler, D. (2012). *Wechsler preschool and primary scale of intelligence—fourth edition*. The Psychological Corporation San Antonio, TX.

Wechsler, D. (2014). Wechsler preschool and primary scale of intelligence—fourth cn edition (wppsi-iv cn).(y. li & z. j, trans. y. li & z. j eds.). *King-May Company China*.

Wu, Z. and Qin, Y. (2012). *Report on Rural Education Development in China 2011*. Beijing Normal University Publishing House.

Wu, Z. and Qin, Y. (2016). *Report on Rural Education Development in China 2015*. Beijing Normal University Publishing House.

Wu, Z. H. (2017). China rural education development report 2017.

Yousafzai, A. K., Obradović, J., Rasheed, M. A., Rizvi, A., Portilla, X. A., Tirado-Strayer, N., Siyal, S., and Memon, U. (2016). Effects of responsive stimulation and nutrition interventions on children's development and growth at age 4 years in a disadvantaged population in pakistan: a longitudinal follow-up of a cluster-randomised factorial effectiveness trial. *The Lancet Global Health*, 4(8):e548–e558.

Yousafzai, A. K., Rasheed, M. A., Rizvi, A., Armstrong, R., and Bhutta, Z. A. (2014). Effect of integrated responsive stimulation and nutrition interventions in the lady health worker programme in pakistan on child development, growth, and health outcomes: a cluster-randomised factorial effectiveness trial. *The Lancet*, 384(9950):1282–1293.

Zhao, L. and Hu, X. (2008). The development of early childhood education in rural areas in china. *Early Years*, 28(2):197–209.

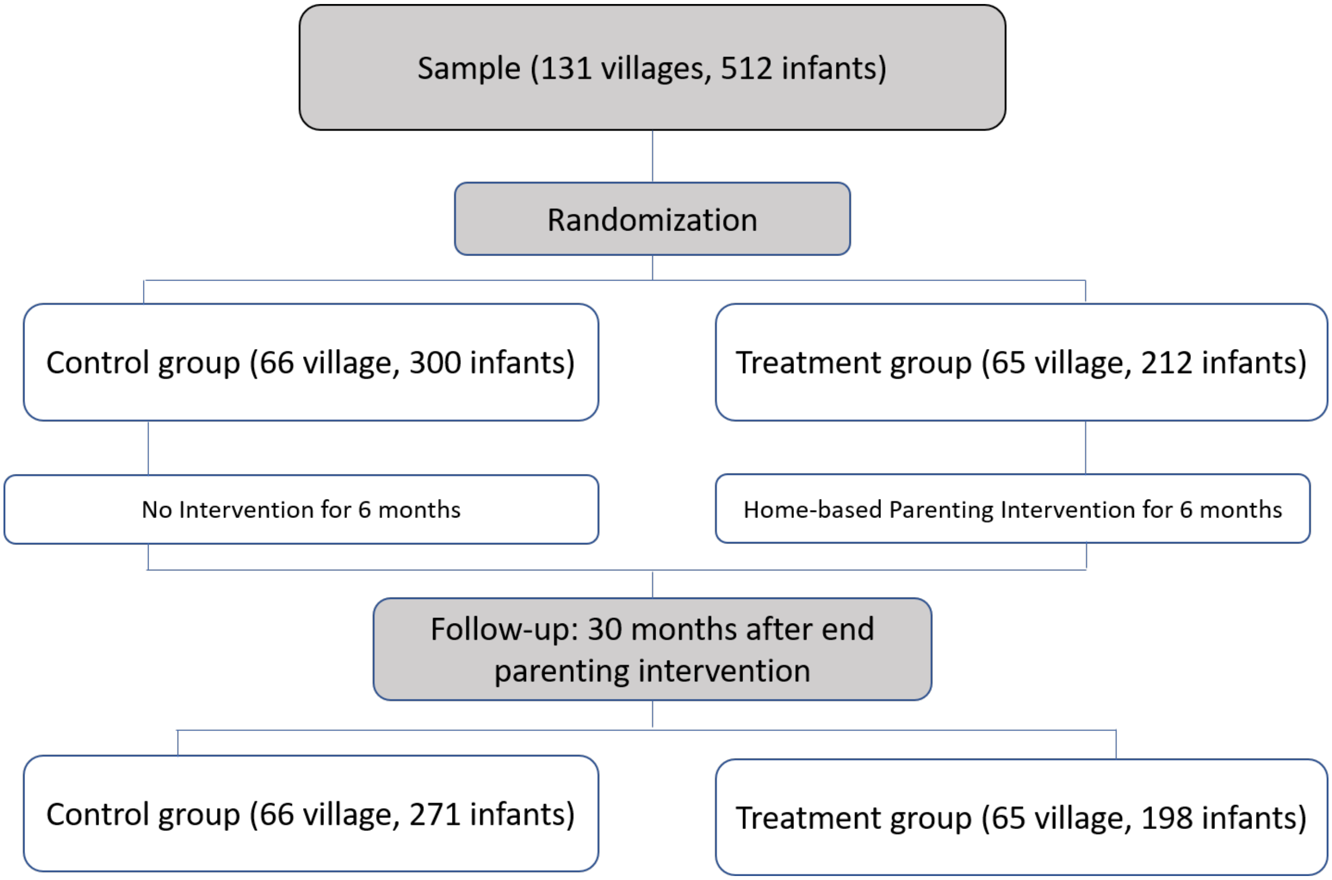
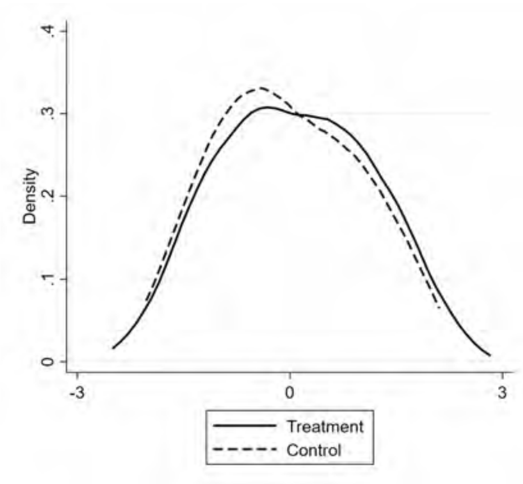
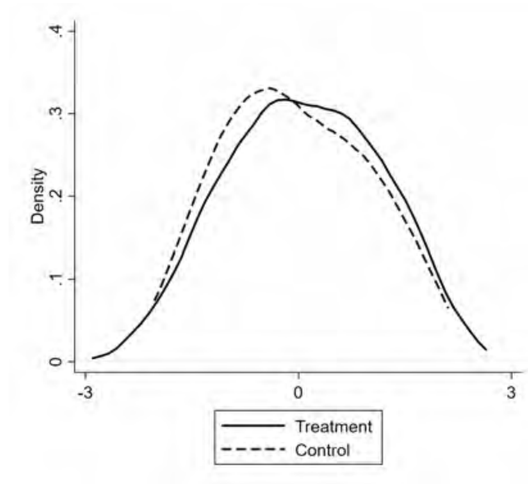


Figure 1: Program Overview

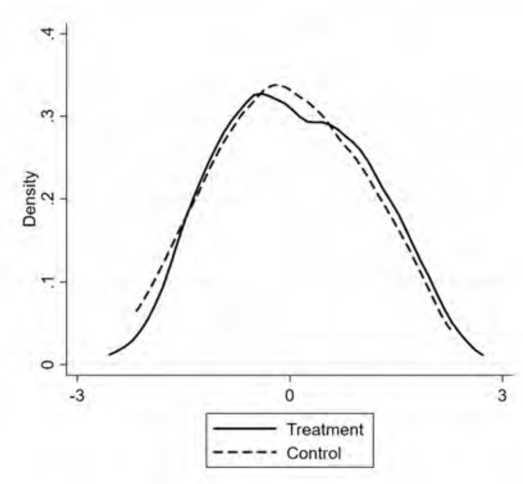
(back to Section 3.3)



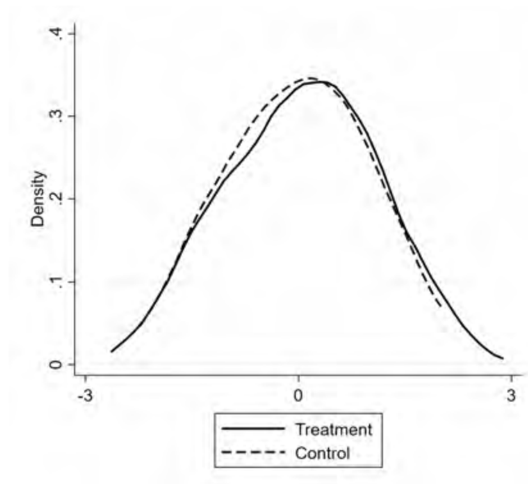
(a) Verbal Comprehension



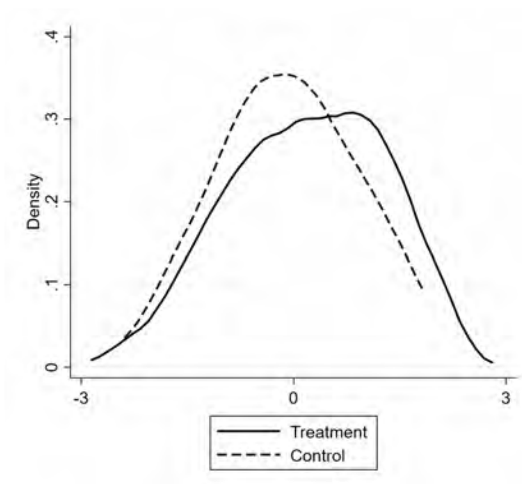
(b) Visual Spatial



(c) Fluid Reasoning

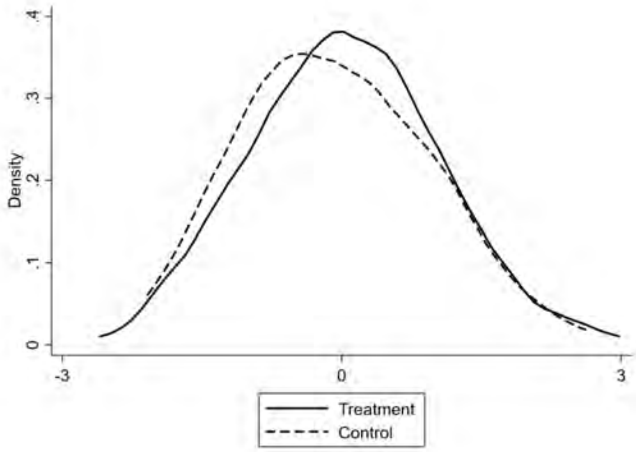


(d) Processing Speed

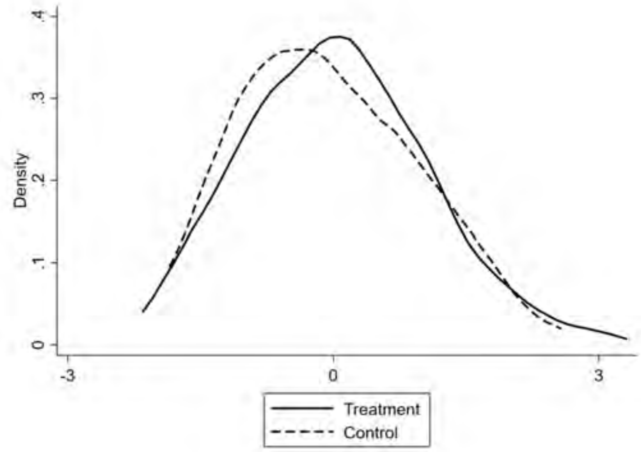


(e) Working Memory

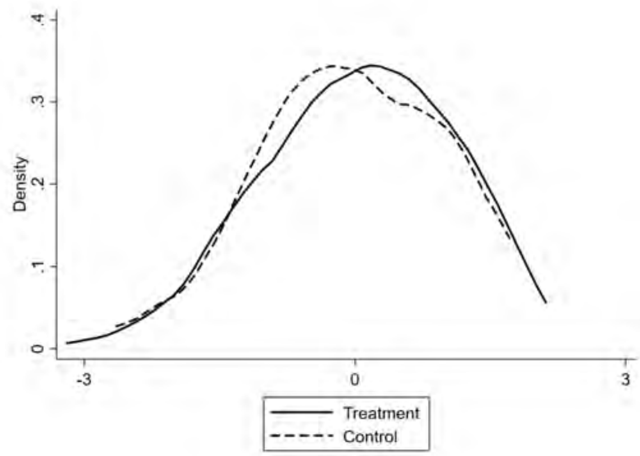
Figure 2: Distribution Cognitive Skill Factors by Treatment Assignment.



(a) Externalizing Behaviour



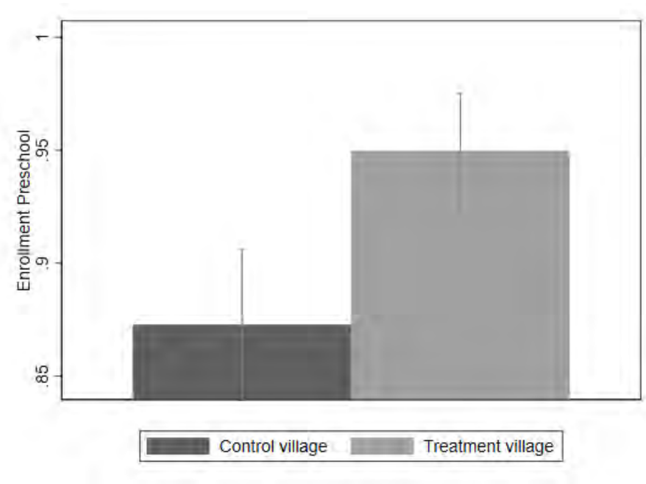
(b) Internalizing Behaviour



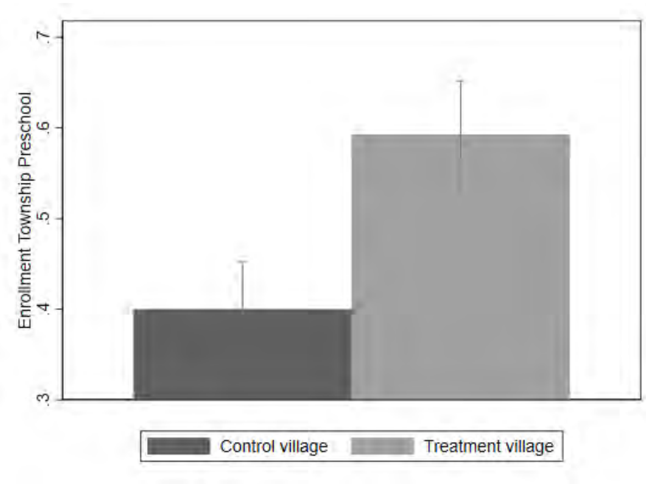
(c) Prosocial Behaviour

Figure 3: Distribution Non-Cognitive Skill Factors by Treatment Assignment

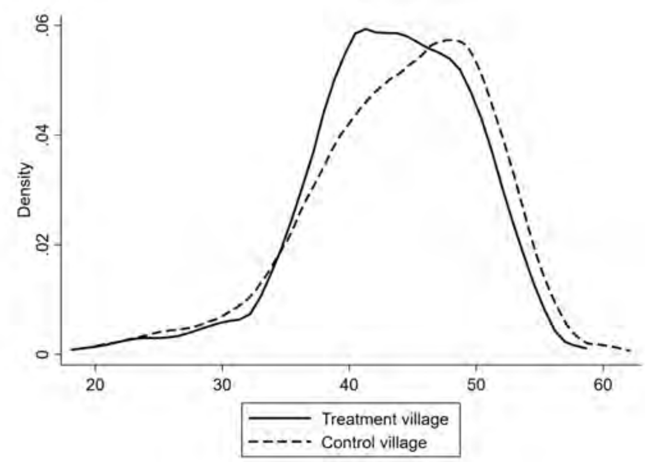
(back to Section 4.1)



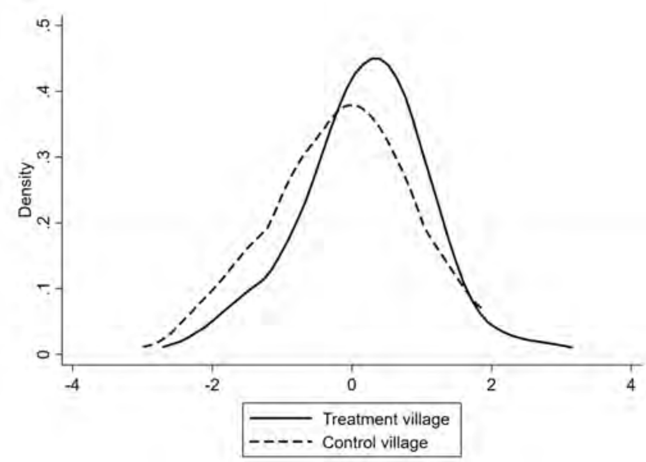
(a) Preschool Enrolment



(b) Preschool Enrolment Township



(c) Preschool Enrolment Age (months)



(d) Preschool Quality

Figure 4: Preschool Enrolment Decision

(back to Section 4.3)

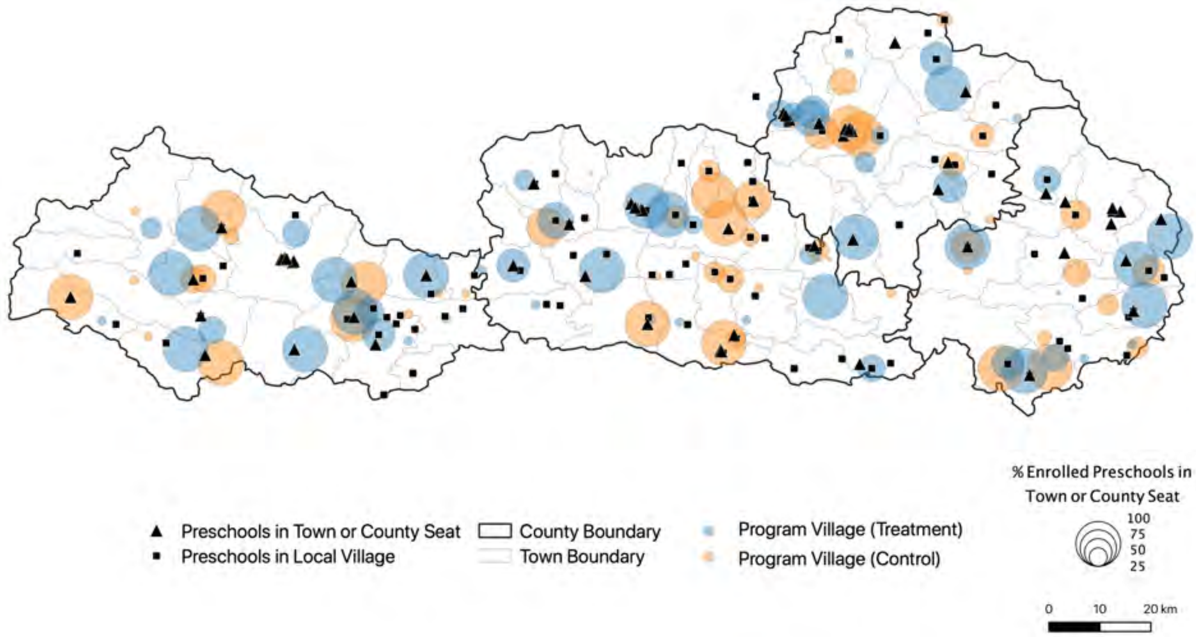


Figure 5: Preschool Enrolment Flows

(back to Section 4.3)

Table 1: Descriptive Statistics and Balance

	(1) Control (N=301)	(2) Treatment (N=212)	(3) p-value
Panel A. Child Characteristics			
(1) Age in months	24.46	24.45	0.747
	(0.20)	(0.22)	
(2) Male	0.45	0.51	0.185
	(0.03)	(0.04)	
(3) Low birth weight	0.04	0.04	0.774
	(0.01)	(0.01)	
(4) First born	0.59	0.61	0.366
	(0.03)	(0.04)	
(5) Ever breastfed	0.85	0.87	0.974
	(0.03)	(0.04)	
(6) Breastfed ≥ 12 months	0.35	0.39	0.867
	(0.05)	(0.05)	
(7) Anemia (Hb <110 g/L)	0.23	0.27	0.849
	(0.03)	(0.04)	
(8) Days ill past month	4.32	4.55	0.620
	(0.33)	(0.37)	
(9) Cognitive Delay (BSID MDI<80)	0.46	0.39	0.206
	(0.04)	(0.03)	
(10) Motor Delay (BSID PDI<80)	0.12	0.10	0.476
	(0.02)	(0.02)	
(11) Social-Emotional Problems (ASQ:SE>60)	0.25	0.28	0.401
	(0.03)	(0.03)	
Panel B. Household Characteristics			
(1) Social security support recipient	0.28	0.25	0.832
	(0.03)	(0.03)	
(2) Mother at home	0.68	0.62	0.116
	(0.04)	(0.05)	
(3) Caregiver education ≥ 9 years	0.72	0.74	0.487
	(0.03)	(0.04)	
(4) Unfavourable perception of FPC	2.87	2.85	0.824
	(0.06)	(0.05)	
Panel C. Parental Inputs			
(1) Told story to child yesterday	0.11	0.11	0.960
	(0.02)	(0.02)	
(2) Read book to child yesterday	0.05	0.04	0.872
	(0.01)	(0.01)	
(3) Sang song to child yesterday	0.37	0.35	0.651
	(0.03)	(0.04)	
(4) Played with child yesterday	0.34	0.34	0.996
	(0.03)	(0.03)	
(5) Number of books in household	1.60	1.90	0.615
	(0.24)	(0.29)	

P-values account for clustering within villages. Unfavourable perception of FPC is measured on a 5-point likert scale.

Table 2: Program Treatment Impact on Infant Cognitive Skills at School-Entry

	Treatment effect			
	Point estimate	Std. error	P-value	FDR q-value
Wechsler Preschool Scale of Intelligence (N=465)				
Verbal Comprehension	0.095	(0.089)	{0.285}	{0.384}
Visual Spatial	0.100	(0.093)	{0.285}	{0.384}
Fluid Reasoning	0.081	(0.086)	{0.347}	{0.384}
Working Memory	0.272***	(0.095)	{0.005}	{0.027}
Processing Speed	0.086	(0.091)	{0.344}	{0.384}

Note: In all regressions we control for strata (county) fixed effects, child gender and baseline developmental outcomes. All skill factors are non-parametrically standardized for each age-month group. To control the potential bias caused by multiple hypothesis testing, we report the rate (q-value) of false discovery rate (FDR)—the proportion of false positives among all positive results. All standard errors are clustered at the village level. Significance levels are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

(back to Section 4.1)

Table 3: Program Treatment Impact on Infant Non-Cognitive Skills at School-Entry

	Treatment effect			
	Point estimate	Std. error	P-value	FDR q-value
Strengths and Difficulties Questionnaire (N=461)				
Externalising behaviour	0.120	(0.080)	{0.136}	{0.617}
Internalising behaviour	0.067	(0.098)	{0.496}	{0.617}
Pro-Social behaviour	0.088	(0.091)	{0.332}	{0.617}

Note: In all regressions we control for strata (county) fixed effects and baseline developmental outcomes. All skill factors are non-parametrically standardized for each age-month group. To control the potential bias caused by multiple hypothesis testing, we report the rate (q-value) of false discovery rate (FDR)—the proportion of false positives among all positive results.All standard errors are clustered at the village level. Significance levels are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

(back to Section 4.1)

Table 4: Program Treatment Impact on Parental Investment at School-Entry

	Treatment effect			
	Point estimate	Std. error	P-value	FDR q-value
Panel A: Investment at Home (N=466)				
Time investment factor	0.298***	(0.108)	{0.007}	{ 0.022}
Material investment factor	0.110	(0.101)	{0.280}	{ 0.107}
Panel B: Investment at Preschool (N= 474)				
Preschool enrolment	0.074**	(0.030)	{0.016}	{ 0.022}
Preschool enrolment age	-1.970**	(1.015)	{0.055}	{0.039}
Preschool quality	0.275**	(0.127)	{0.030}	{0.025}

Note: In all regressions we control for strata (county) fixed effects and baseline developmental outcomes. Around 10% of children are not yet enrolled at the time of the survey hence we assume they will enrol by the maximum age in the sample distribution. Program treatment effects on township preschool enrollment, preschool enrollment age, and preschool quality selection are conditional on being currently enrolled (N=429). Preschool information is missing for 22 households. We impute the preschool quality index for those missing observations with the township average preschool quality by the locations (township or village) of of preschools that households enroll the child. To control the potential bias caused by multiple hypothesis testing, we report the rate (q-value) of false discovery rate (FDR)—the proportion of false positives among all positive results. All standard errors are clustered at the village level. Significance levels are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

(back to Section 4.2)

Table 5: Program Treatment Impact on Preschool Location Selection

	Treatment effect			
	Point estimate	Std. error	P-value	FDR q-value
Preschool Location Selection (N=429)				
Enrol at township preschool (1=yes)	0.184***	(0.066)	{0.006}	{0.024}
Enrol at closest preschool (1=yes)	-0.182**	(0.070)	{0.011}	{0.024}

Note: In all regressions we control for strata (county) fixed effects and baseline developmental outcomes. Around 10% of children are not yet enrolled at the time of the survey hence we assume they will enrol by the maximum age in the sample distribution. Preschool information is missing for 22 households. Enrol at closest preschool variable is a binary variable that equals one when the distance to the enrolled preschool is larger than the distance to the closet preschool. To control the potential bias caused by multiple hypothesis testing, we report the rate (q-value) of false discovery rate (FDR)—the proportion of false positives among all positive results. All standard errors are clustered at the village level. Significance levels are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

(back to Section 4.3)

Table 6: Parents' Preferences and Elasticity of Demand with respect to Preschool Attributes

	Reference Group	Interaction	Initial share	Mean Preschool	Elasticity of Demand	
	(Control Group)	(Treatment Group)	of Q	Characteristics	Control Group	Treatment Group
Total Preschool Expense (thousand)	-0.070	0.170	0.03	1.17	-0.079	0.114
(Tuition+other fees)	(0.084)	(0.147)			[0.098]	[0.159]
Preschool Quality (SD)	0.056	0.384**	0.03	3.11	0.171	1.333***
	(0.090)	(0.155)			[0.279]	[0.386]
Distance rank	-0.278***	-0.059	0.03	18.65	-5.035***	-6.094***
	(0.029)	(0.052)			[0.591]	[0.869]
Observations			13934			

Note: Standard errors clustered at household level are in parentheses. Distance rank is used in the estimation instead of actual distance to limit the influence of the outliers. The elasticity of demand for each group is calculated using the formula $\frac{\partial Q}{\partial X} \times \frac{X}{Q}$, which for the reference group (Control Group) is $\beta \times X \times (1 - Q)$ and for the treatment group (denoted by t) is $(\beta + \beta_t) \times X \times (1 - Q)$. We compute the elasticity with respect to each continuous school attribute X using the mean of X in the same sample. The distribution of the school quality were shifted so that all observations have positive value. The mean number of schools in the household's choice is 33.3. We therefore fix Q to be $1/33.3 \approx 0.03$. Bootstrapped standard errors (based on 200 repetitions) are reported in brackets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

(back to Section 4.3)

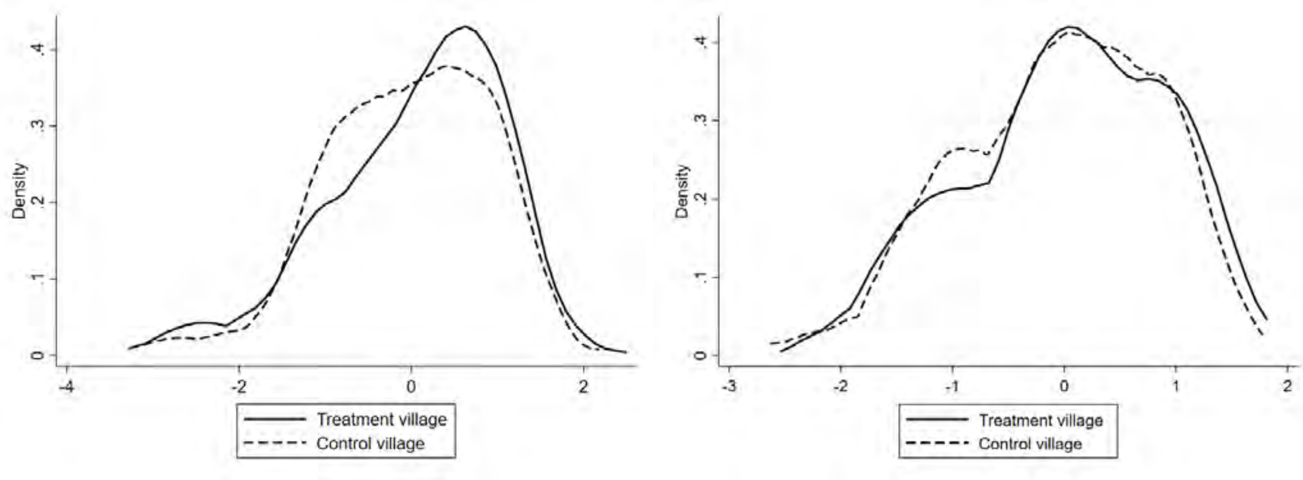
Table 7: Mediation Analysis of Program Impact on Infant Working Memory Skill

	(1)	(2)	(3)	(4)	(5)
	Working Memory Infant Skill				
Treatment	0.252***	0.256***	0.243**	0.212**	0.213**
	(0.093)	(0.094)	(0.094)	(0.092)	(0.094)
		[0.24, 0.256]	[0.205, 0.243]	[0.124, 0.212]	[0.125, 0.213]
Time investment factor		-0.012			-0.015
		(0.044)			(0.044)
		[-0.149, -0.012]			
Material investment factor			0.061*		0.041
			(0.034)		(0.034)
			[-0.043, 0.061]		
Preschool quality factor				0.122**	0.113**
				(0.052)	(0.053)
				[0.058, 0.122]	
Observations	458	458	458	458	458
R^2	0.18	0.18	0.18	0.19	0.19
Oster's $\delta_{(mediator)}$		-0.107	0.622	1.627	

Note: In all regressions we control for strata (county) fixed effects, gender of the child, number of siblings, baseline developmental outcomes, caregiver completed at least 9 years of education, baseline parenting time investment factor, village average of baseline development outcomes, and township level characteristics, including population, a average income, labor force participation and labor force out-migration. Identified set using Oster (2019) is reported in brackets, given R^2_{max} equals 0.403. Around 10% of children are not yet enrolled at the time of the survey, hence in the mediation analysis through preschool quality factor, the preschool quality of not enrolled children is coded with the mean preschool quality of the township. Significance levels are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

(back to Section 4.4)

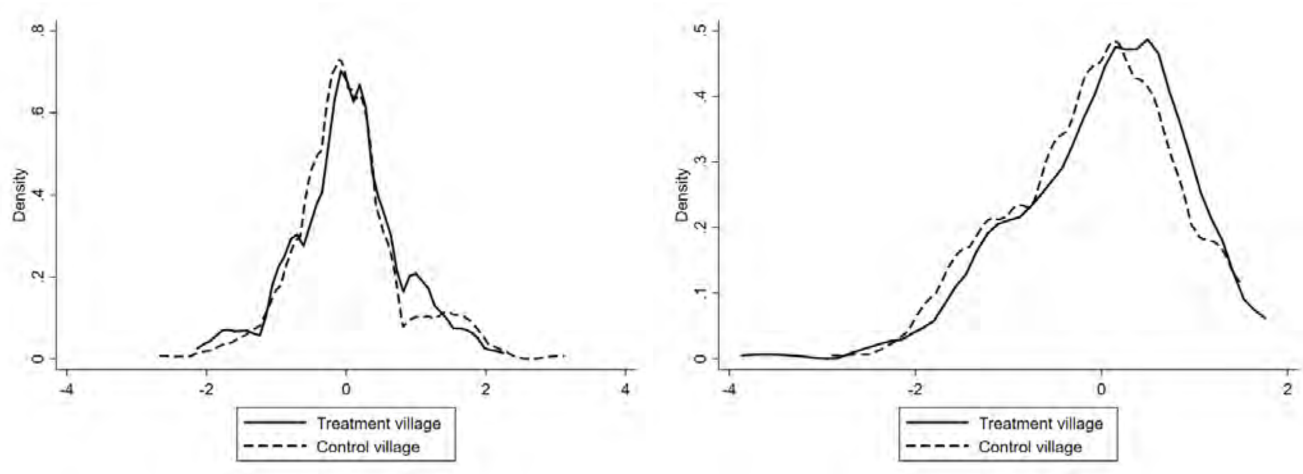
Appendix A



(a) Information Test

(b) Similarities Test

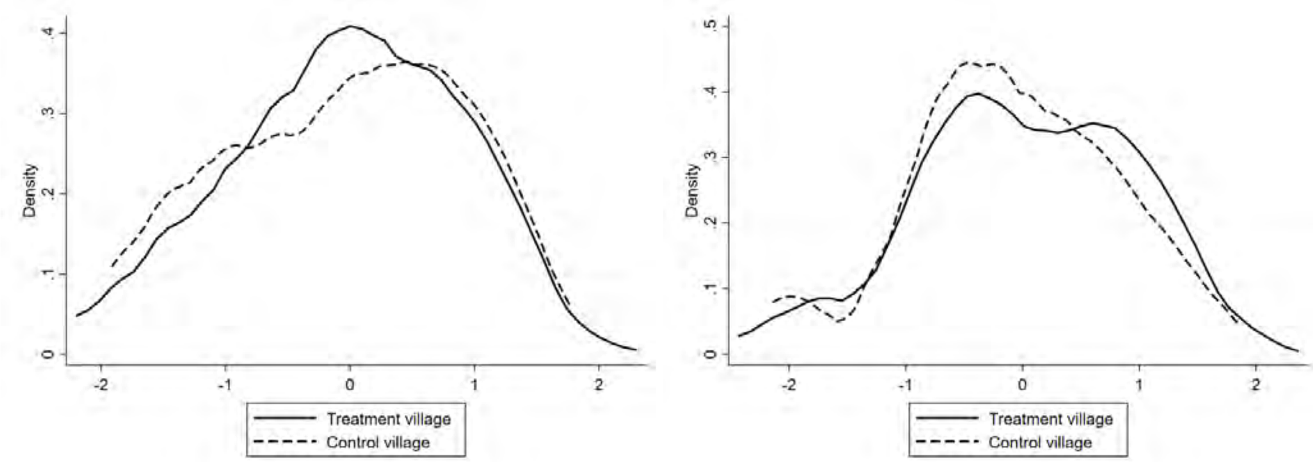
Figure A1: Distribution WPPSI-IV Verbal Comprehension Skills by Treatment Assignment



(a) Block Design Test

(b) Object Assembly Test

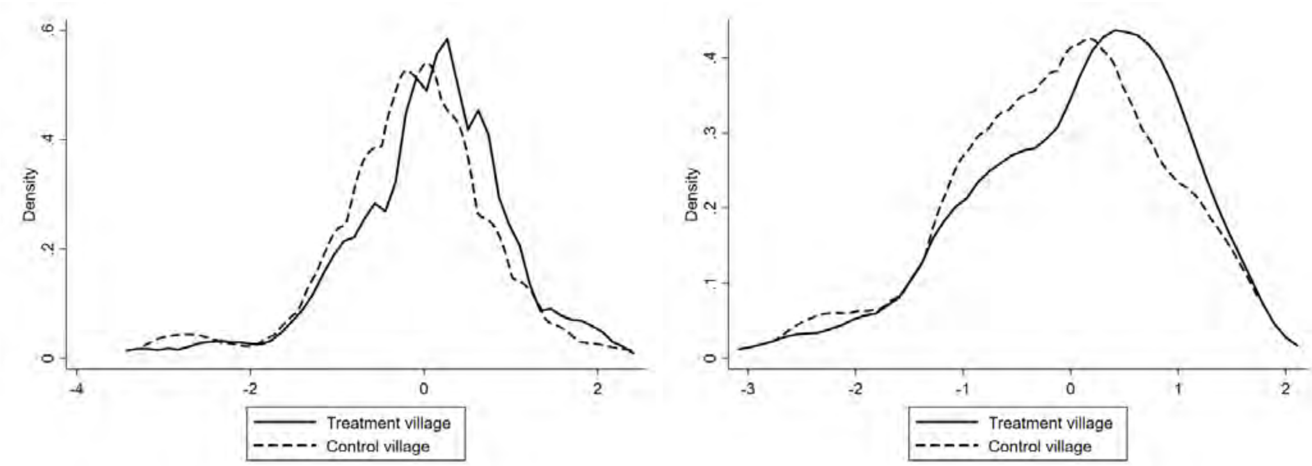
Figure A2: Distribution WPPSI-IV Visual Spatial Skills by Treatment Assignment



(a) Matrix Reasoning Test

(b) Picture Concepts Test

Figure A3: Distribution WPPSI-IV Fluid Reasoning Skills by Treatment Assignment



(a) Zoo Locations Test

(b) Picture Memory Test

Figure A4: Distribution WPPSI-IV Working Memory Skills by Treatment Assignment

(back to Section 3.3)

Table A1: Descriptive Statistics and Balance Adjusted for Attrition

	(1) Control (N=276)	(2) Treatment (N=199)	(3) P-value
Panel A. Child Characteristics			
(1) Age in months	25.019	25.075	0.9514
	[3.284]	[3.367]	
(2) Male	0.454	0.525	0.1241
	[0.499]	[0.501]	
(3) Low birth weight	0.038	0.041	0.9411
	[0.191]	[0.199]	
(4) First born	0.568	0.601	0.3566
	[0.496]	[0.491]	
(5) Anemia (Hb <110 g/L)	0.241	0.283	0.7963
	[0.428]	[0.452]	
(6) Days ill past month	4.459	4.589	0.8192
	[5.150]	[5.463]	
(7) Cognitive Delay (BSID MDI<80)	0.457	0.389	0.3045
	[0.499]	[0.489]	
(8) Motor Delay (BSID PDI<80)	0.113	0.102	0.7960
	[0.318]	[0.303]	
(9) Social-Emotional Problems(ASQ:SE>60)	0.255	0.283	0.5454
	[0.437]	[0.452]	
Panel B. Household Characteristics			
(1) Social security support recipient	0.351	0.333	0.8354
	[0.478]	[0.473]	
(20 Mother at home	0.684	0.636	0.2079
	[0.466]	[0.482]	
(3) Caregiver education ≥ 9 years	0.652	0.657	0.7494
	[0.477]	[0.476]	
(4) Unfavourable perception of FPC	2.832	2.832	0.9460
	[0.600]	[0.630]	
Panel C. Parental Inputs			
(1) Told story to child yesterday	0.115	0.117	0.9338
	[0.319]	[0.322]	
(2) Read book to child yesterday	0.045	0.041	0.7929
	[0.208]	[0.198]	
(3) Sang song to child yesterday	0.373	0.354	0.6359
	[0.485]	[0.479]	
(4) Played with child yesterday	0.324	0.354	0.5966
	[0.469]	[0.479]	
(5) Number of books in household	1.512	1.924	0.4193
	[3.475]	[4.644]	

Standard deviation in the bracket; P-values account for clustering within villages.

Table A2a: Preschool and Teacher Characteristics

	(1) Village	(2) Township	(3) p-value
Number of pupils	69.585	251.087	0.000
	(8.777)	(19.146)	
Share pupils receiving government need-based aid	0.268	0.208	0.043
	(0.021)	(0.018)	
Tuition fee per semester (Yuan)	607.926	1275.704	0.001
	(25.942)	(226.249)	
Teacher age	36.128	33.377	0.061
	(1.041)	(0.936)	
Teacher male	0.138	0.000	0.001
	(0.036)	(0.000)	
Teacher experience	6.452	4.819	0.119
	(0.814)	(0.498)	
Teacher monthly salary (Yuan)	2183.436	2158.086	0.892
	(136.098)	(114.954)	
Share of teachers with bachelelor degree	0.251	0.528	0.000
	(0.036)	(0.042)	
Teacher training in past year	0.691	1.000	0.000
	(0.048)	(0.000)	
N	94	71	

Standard errors in parentheses.

Table A2b: Preschool Quality

	(1) Village	(2) Township	(3) p-value
Panel A: Structural Quality			
Pupil-teacher ratio	21.964	20.169	0.406
	(1.620)	(1.208)	
Number of activity rooms	1.862	6.333	0.000
	(0.216)	(0.454)	
Outdoor play-area	368.426	668.913	0.017
	(77.270)	(100.475)	
Preschool has play room	0.160	0.406	0.000
	(0.038)	(0.060)	
Preschool has exercise room	0.149	0.464	0.000
	(0.037)	(0.060)	
Preschool has dormitories	0.213	0.609	0.000
	(0.042)	(0.059)	
Preschool provides breakfast	0.500	0.696	0.012
	(0.052)	(0.056)	
Panel B: Process Quality			
Teacher reads books in class	0.926	1.000	0.020
	(0.027)	(0.000)	
Teacher organizes exercise activities	0.702	0.971	0.000
	(0.047)	(0.020)	
Teacher organizes art& music activities	0.936	0.971	0.312
	(0.025)	(0.020)	
Teacher organizes science activities	0.809	0.942	0.014
	(0.041)	(0.028)	
Teacher teaches social skills	0.957	0.971	0.652
	(0.021)	(0.020)	
Teacher teaches language skills	0.989	0.971	0.392
	(0.011)	(0.020)	
N	94	71	

Standard errors in parentheses.

(back to Section 4.3)

Appendix B: Measurement System

This appendix provides further detail about the dimensionality reduction techniques and measurement systems of infant skills, parental investment and school quality.

B.1. Infant Cognitive Skills

Infant cognitive skills are measured using the Wechsler Preschool and Primary Scale of Intelligence (WPPSI-IV)(Wechsler, 2012). The WPPSI-IV consists of several individually administered subtests, each of which measures a specific area of cognitive ability. In each sub-test in index (i)-(iv) administered test-items increase in difficulty level and the test is stopped when a child can no longer provide a correct answer. Given this specific test structure, we first estimate a two-parameter logistic IRT measurement system for each of the eight sub-tests which calculates the optimal weighted average of all items taking into account response patters. Conceptually, IRT models can be viewed as an extension of confirmatory factor analysis (CFA) to binary or categorical outcomes.[22]

Let I_{ij}^{λ} define the performance measure for child i and item j on test λ and let's assume it is determined as follows:

$$I_{ij}^{\lambda} = \beta_j + \alpha_j \Lambda_i^{\lambda} + \epsilon_{ij}^{\lambda} \tag{5}$$

where Λ_i^{λ} is child i's latent skill for test λ and this is assumed to be independent from the error term ϵ_{ij}^{λ}. In other words, we assume that a unidimensional skill is sufficient in explaining a child's response behaviour on items in each sub-test. We further assume that a child's response to an item is independent of his or her responses to other items after conditioning on child latent skill.

The variable I_{ij}^{λ} is not observed to the enumerator or caregiver. Instead, we observe I_{ij}^{λ}=1 if $I_{ij}^{\lambda} > 0$ and $I_{ij}^{\lambda} = 0$ otherwise. The model is identified by assuming $\Lambda_i^{\lambda} \sim N(0,1)$. We further assume that the measurement system is invariant to treatment assignment. Hence, the probability

[22]See Skrondal and Rabe-Hesketh (2009) for detailed overview of IRT estimation methods and Zhao and Hu (2008) for practical examples.

of observing $I_{ij}^{\lambda} = 1$ given the child's latent skill Λ_i^{λ} is denoted as follows:

$$Pr(I_{ij}^{\lambda} = 1|\Lambda_i^{\lambda}) = 1 - Pr(\epsilon_{ij}^{\lambda} \leq -\beta_j - \alpha_j \Lambda_i^{\lambda}|\Lambda_i^{\lambda}) = 1 - F_{\epsilon}(-\beta_j - \alpha_j \Lambda_i^{\lambda}) \tag{6}$$

where F_{ϵ} is the distribution of ϵ_{ij}^{λ}. The estimated item specific intercepts, β_j, represent the level of difficulty of item j. The higher the value of β_j, the lower the success rate of item j is for a given latent skill level and hence the more difficult item j is. The parameter α_j represents the discrimination ability of item j as the rate at which the probability of answering correctly changes with a child's latent skill. Items with large discrimination value have a high correlation between latent skill and the probability of success and can distinguish better between low and high levels of latent skill. Hence, in the 2-parameter logistic IRT model, the probability of success on an item j is a function of both the level of latent skill Λ_i^{λ} and the difficulty level, β_j, and discrimination ability, α_j, of item j. Below we describe the individual tests in more detail and provide test diagnostics from the two-parameter logistic IRT model.

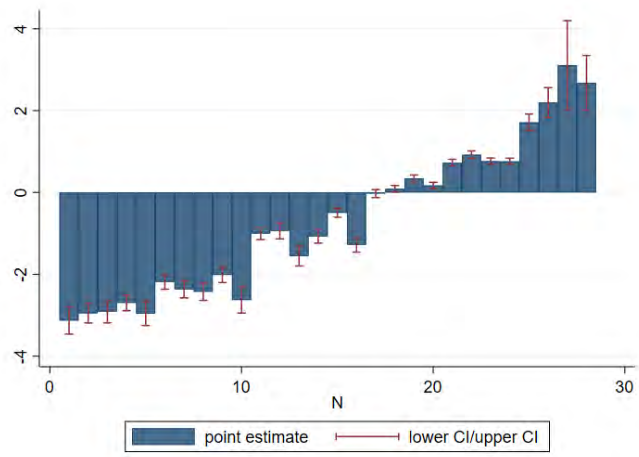
(i) Verbal Comprehension

The verbal comprehension index measures a child's verbal reasoning and comprehension abilities and is assessed using the WPPSI-IV Similarities test and the Information test. During the administration of the Similarities test the child is read incomplete sentences containing two concepts that share a common characteristic. The child is asked to complete the sentence by providing a response that reflects the shared characteristic. During the Information test, the child is asked to respond to questions by choosing pictures from four response options and answer questions addressing a broad range of general knowledge topics.

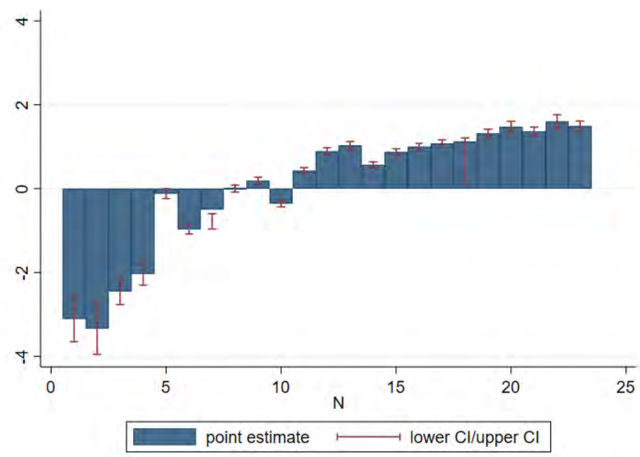
Figure B1.1 plots the distribution of estimated item-difficulty parameters , $\hat{\beta}_j$, for each of the 2 administered tests from the two-parameter logistic IRT measurement system. Items with negative estimated difficulty parameters are considered relatively easy, and items with positive difficulty parameters are relatively hard. The distribution of the estimated difficulty parameters provides information about whether the test is well-designed for the population under study.

In an ideal test, the difficulty parameters smoothly transition from easy to more difficult and cover the whole skill distribution. By this metric, the Similarities test is not well designed as the values of the difficulty level are relatively flat after item 11. This means that the test is not able to distinguish well between children with medium and high latent skill levels.

Figure B1.2 plots the Item Characteristic Curves (ICCs) for both tests. The ICC plots the probability that a person is successful on a given item j as a function of child's i latent skill Λ_i^λ for test λ. The ICCs for the tests of the WPPSI-IV Verbal Comprehension index confirm that the Similarities test fails to differentiate between medium and higher levels of latent skill. Moreover, the estimated discrimination parameters, $\hat{\alpha}_j$, for more easy items are relatively low indicating that the test is also not very good in differentiating between low and medium latent skill.

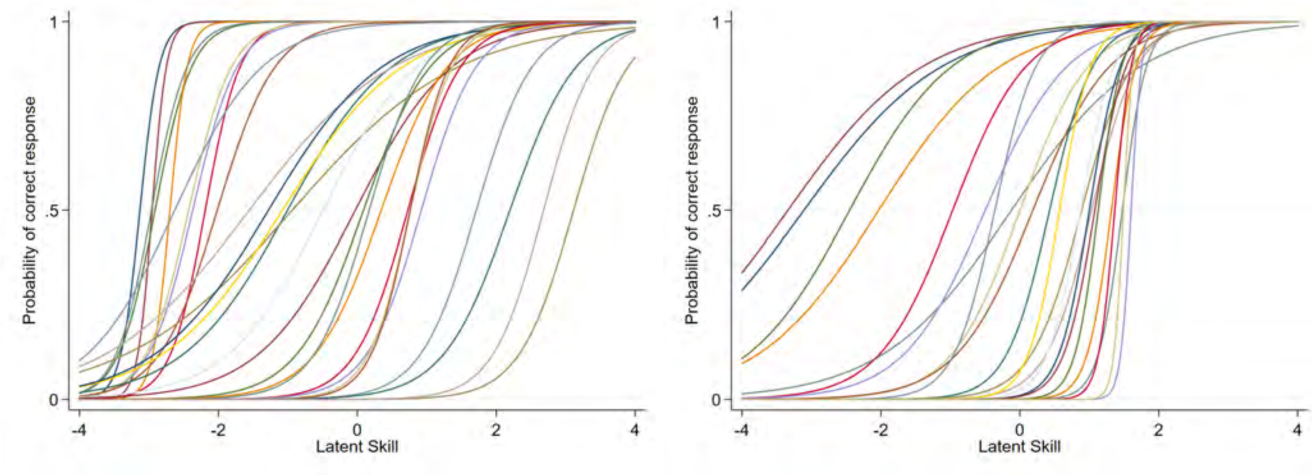


(a) Information Test



(b) Similarities Test

Figure B1.1: Distribution of task item difficulty levels WPPSI-IV Verbal Comprehension



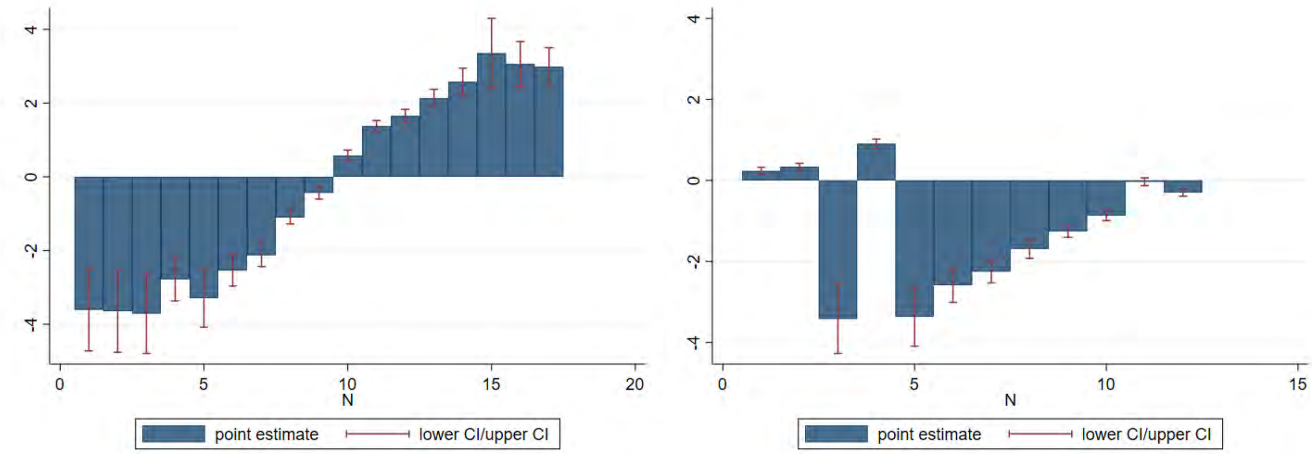
(a) Information Test

(b) Similarities Test

Figure B1.2: Item Characteristic Curves (ICCs) WPPSI-IV Verbal Comprehension

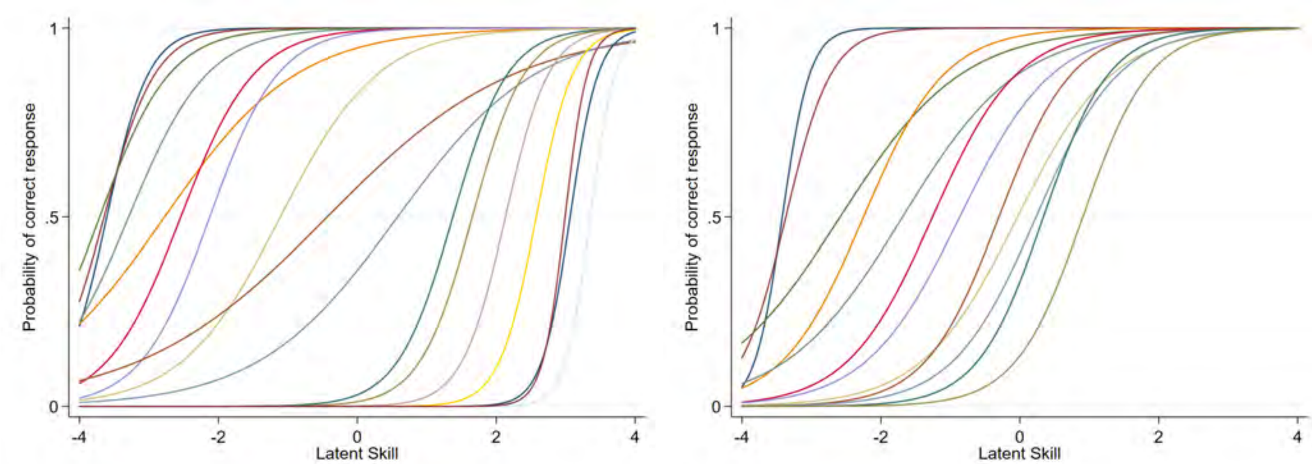
(ii) Visual Spatial

The visual spatial index measures the ability to organise and understand visual parts and information, assimilate visual and motor functions simultaneously, and see the whole-part connection to objects. Visual spatial ability is assessed using the WPPSI-IV Block Design and Object Assembly test. During the Block design test, a child is asked to use one- or two-color blocks to recreate the design of a picture in a stimulus book within a specific time limit. In the Object Assembly test, the child is presented with pieces of a puzzle which needs to be fit together within the time span of 90 seconds. Figures B2.1- B2.2. show that the Block Design test is good in distinguishing between low and high ability, but less accurate in measuring medium ability. The Object Assembly test, on the other hand, is only precise in measuring very low latent skill as almost all estimated item difficulty parameters are negative.



(a) Block Design Test (b) Object Assembly Test

Figure B2.1: Distribution of task item difficulty levels WPPSI-IV Visual Spatial

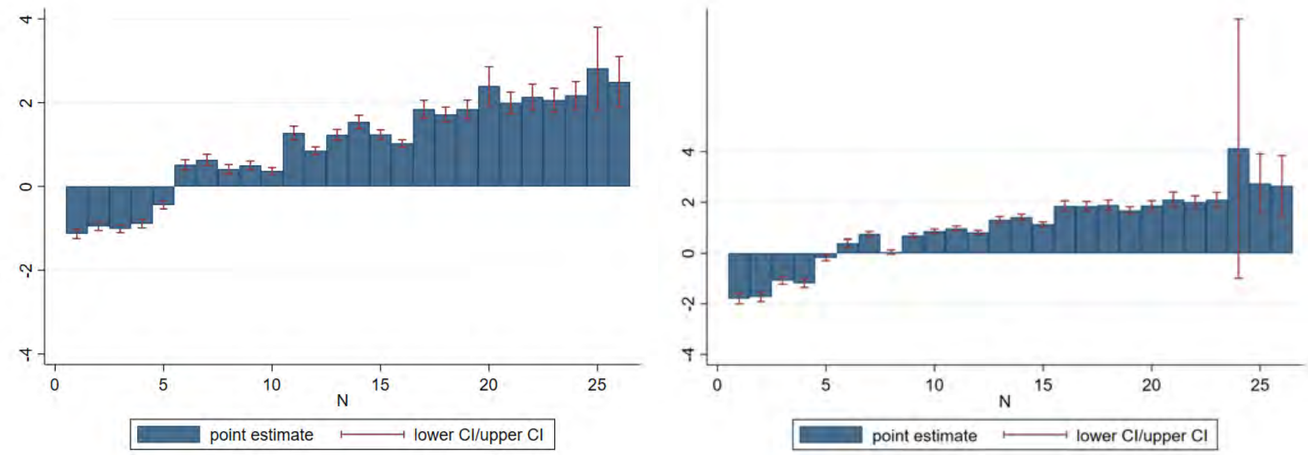


(a) Block Design Test (b) Object Assembly Test

Figure B2.2: Item Characteristic Curves (ICCs) WPPSI-IV Visual Spatial

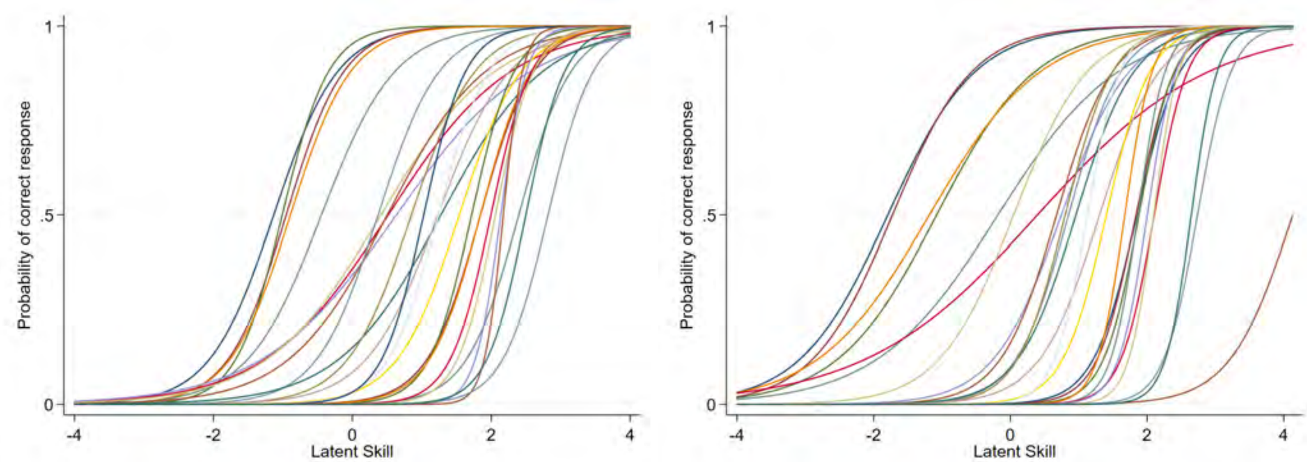
(iii) Fluid Reasoning

The fluid reasoning index measures a child's ability to utilize inductive reasoning which is the ability to use past observations to predict current situations. Fluid reasoning ability is administered by the WPPSI-IV Matrix Reasoning and Picture Concept test. In the Matrix Reasoning test, a child is presented with an incomplete matrix and asked to select missing parts from 4 or 5 response options. During the Picture Concepts test, the child is shown two or three rows of pictures and needs to choose one picture from each row to form a group with a common characteristic. Figures B3.1- B3.2 show that both tests are relatively better at measuring and distinguishing between higher latent skill levels. Figures B3.1- B3.2 show that both tests are relatively better at measuring and distinguishing between higher latent skill levels.



(a) Matrix Reasoning Test (b) Picture Concepts Test

Figure B3.1: Distribution of task item difficulty levels WPPSI-IV Fluid Reasoning

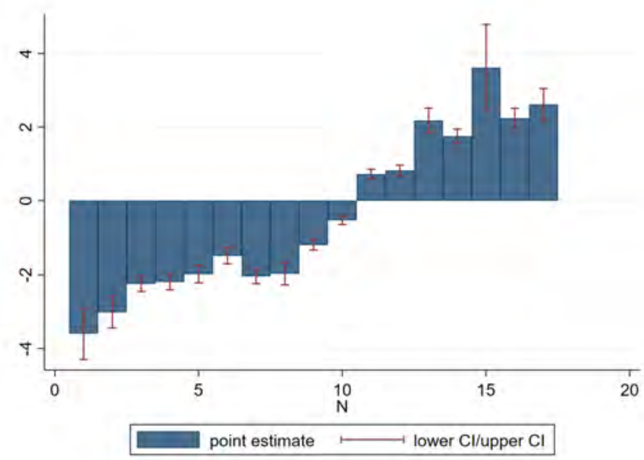


(a) Matrix Reasoning Test (b) Picture Concepts Test

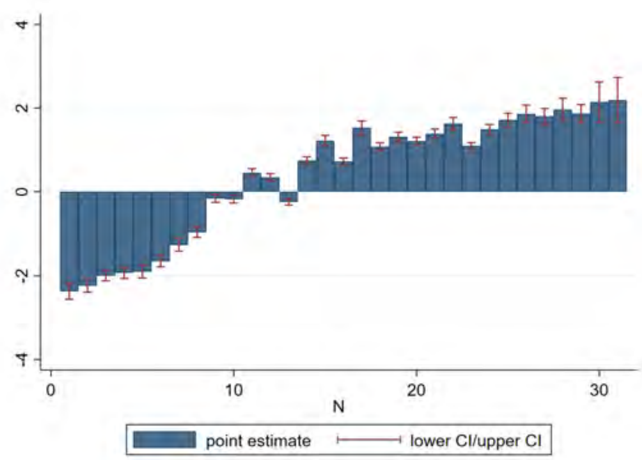
Figure B3.2: Item Characteristic Curves (ICCs) WPPSI-IV Fluid Reasoning

(iv) Working Memory

The working memory index measures the ability to balance focus and attention while manipulating visual and auditory information in conscious awareness and is administered by the WPPSI-IV Zoo Location and Picture Memory tests. In the Zoo Locations test, a child is shown one or more animal card placeds on a zoo layout and than asked to place the animal cards in the previously displayed locations. During the Picture Memory test, a child is shown one or more pictures for a specific duration of time and than asked to select the same picture from options on a response page. Figures B4.1- B4.2 show that the Zoo Locations test is relatively good in distinguishing between different latent skill levels but has a small amount of items which reduces overall accuracy. The Picture Memory test is well designed to distinguish between low, medium and high latent skills and has high discrimination ability across all items.

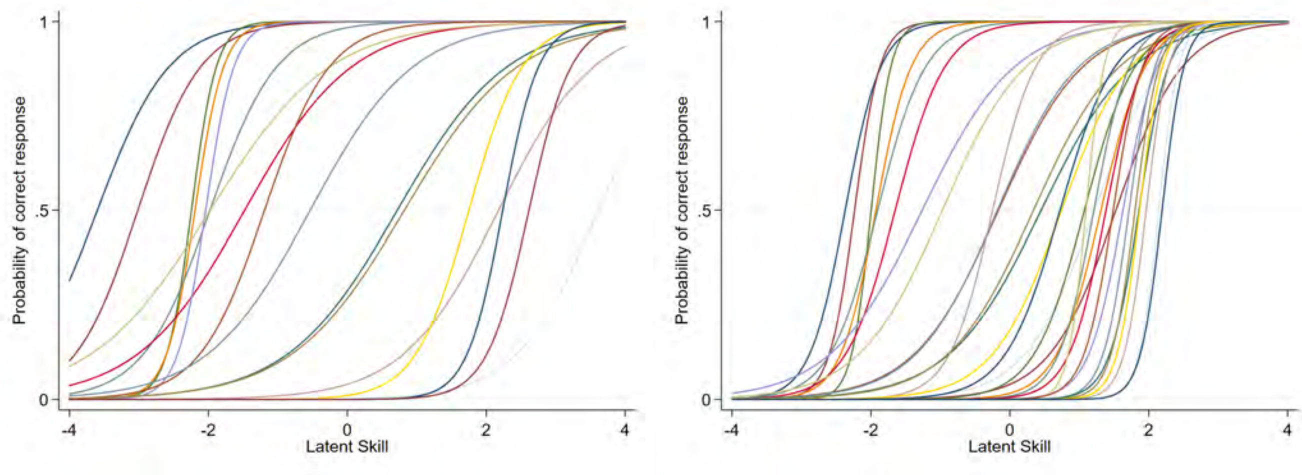


(a) Zoo Locations Test



(b) Picture Memory Test

Figure B4.1: Distribution of task item difficulty levels WPPSI-IV Working Memory



(a) Zoo Locations Test (b) Picture Memory Test

Figure B4.2: Item Characteristic Curves (ICCs) WPPSI-IV Working Memory

(v) Processing Speed

The processing speed index analyses how quickly a child can scan and differentiate visual information and is administered by the WPPSI-IV Bug Search and Cancellation tests. Both tests measure the amount of time a child requires to finish the task. During the Bug Search test, a child uses an ink dauber to mark the image of a bug that matches the target bug in a collection of different bugs. For the Cancellation test, a child is asked to scan two arrangements of objects on a page and mark all the target objects. Performance on both test is not dependent on increasing task difficulty as is the case for the other WPPSI-IV indexes and hence there is no need to estimate an IRT model.

B.2. Infant Non-Cognitive Skills

Infant non-cognitive skills are measured using the Strengths and Difficulty Questionnaire (SDQ), a widely used behavioural screening tool translated to Chinese and validated on a Chinese sample (Goodman et al., 2000; Du et al., 2008). The SDQ comprises of 25 items assessing social, emotional and behavioural functioning of children reported by the main caregiver on a 3-point likert scale (1 *not true*, 2 *somewhat true*, 3 *certainly true*). Items are both positively and negatively phrased to avoid the effect of acquiescence bias. The original proposed factor structure of the

SDQ includes five scales of five items each corresponding with five sub-domains: (i) conduct problems; (ii) hyperactivity/inattention; (iii) emotional symptoms; (iv) peer problems and (v) prosocial behaviour (Goodman, 1997).

Alternatively sub-domains (i) and (ii) can be combined to measure *externalizing behaviour* which assesses behavioural problems that are manifested in children's outward behaviour such as disruptiveness, hyperactivity, and aggressive behaviour. The sub-domains (iii)-(iv) can be combined to measure *internalizing behaviour* which assesses behavioural problems affecting children's internal psychological environment such as withdrawn, anxious, and depressed behaviour. More recent studies using exploratory and confirmatory factor analysis indicate that this three-factor structure might be more appropriate (Dickey and Blumberg, 2004; Goodman et al., 2010).

We follow the literature and estimate a three-factor dedicated measurement system in which each item is associated with at most one factor (Gorsuch, 2003; Thompson, 2004). Parameters of the dedicated measurement system are estimated using maximum likelihood and can be found in Table B1. The first column in Table B1 reports factor loadings of each of the 25 items for the three non-cognitive skill factors. We normalize factor loadings of the first measure for each skill factor to one. In the second column of Table B1 we report the signal-to-noise ratio which indicates how much of the variance in each of the 25 items is driven by signal relative to noise. The signal-to-noise ratios for the j^{th} item is calculated as:

$$S_j = \frac{\gamma_j^2 Var(\theta^\lambda)}{\gamma_j^2 Var(\theta^\lambda) + Var(\delta_j^\lambda)}$$

Several items of the SDQ have poor signal-to-noise ratios, confirming previous findings in the literature that document measurement error in early childhood skills (Cunha et al., 2010), especially for caregiver assessments as correlations between questions are partially driven by answering patterns of respondents (Johnston et al., 2014; Laajaj and Macours, 2017).

Table B1: Measurement System Non-Cognitive Skills

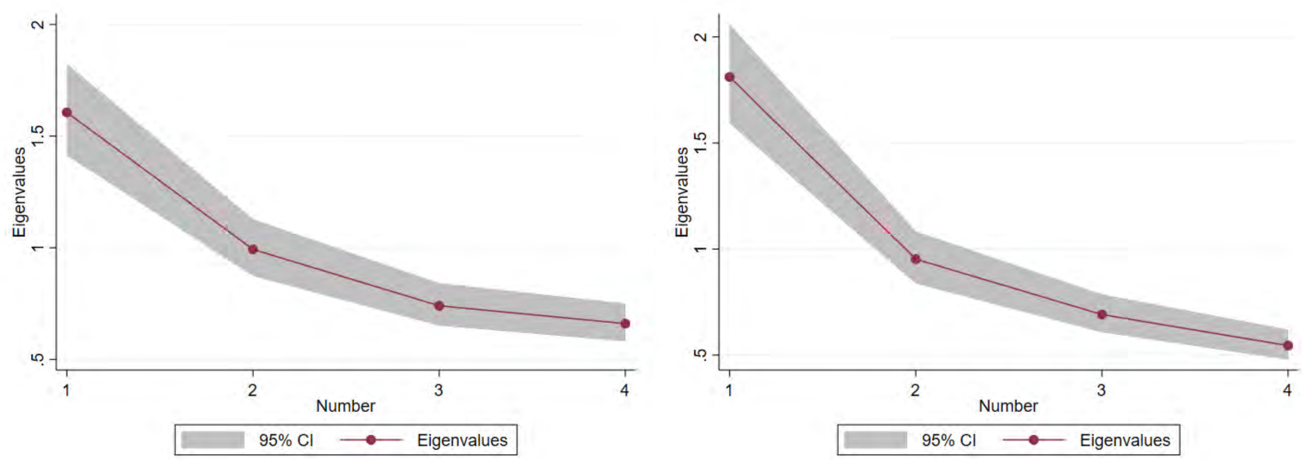
Latent skill factor	Measurement item	Factor loading	% signal
Externalizing Behaviour	Often loses temper	1	0.317
	Generally well behaved, usually does what adults request	-0.220	0.014
	Often fights with other children or bullies them	0.485	0.151
	Often lies or cheats	0.461	0.145
	Steals from home, school or elsewhere	0.121	0.060
	Restless, overactive, cannot stay still for long	0.728	0.139
	Constantly fidgeting or squirming	1.014	0.247
	Easily distracted, concentration wanders	0.686	0.177
	Think things out before acting	-0.291	0.030
	Good attention span, sees work through to the end	-0.503	0.081
Internalizing Behaviour	Often complains of headaches, stomach-aches or sickness	1	0.174
	Many worries or often seems worried	1.000	0.215
	Often unhappy, depressed or tearful	1.031	0.209
	Nervous or clingy in new situations, easily loses confidence	1.084	0.150
	Many fears, easily scared	1.344	0.271
	Rather solitary, prefers to play alone	0.747	0.090
	Has at least one good friend	-0.135	0.005
	Generally liked by other children	-0.244	0.013
	Picked on or bullied by other children	0.947	0.181
	Gets along better with adults than with other children	0.247	0.008
Prosocial Behaviour	Considerate of other people's feelings	1	0.124
	Shares readily with other children, for example toys, treats,	0.998	0.145
	Helpful if someone is hurt, upset or feeling ill	1.722	0.334
	Kind to younger children	1.108	0.208
	Often offers to help others (parents, teachers, or other children)	1.697	0.411

B.3. Parental Investment

In a first step we use exploratory factor analysis (EFA) to determine the number of factors we need to extract from our list of time and material parental investment measures. We use Horn's parallel analysis (Horn, 1965) and Cattell's scree plot (Cattell, 1966) to guide us in the factor selection process (Table B2 and Figure B5).

Table B2: EFA Factor Selection Parental Investment

	Cattell's scree plot	Horn's parallel analysis
Time investment	1	1
Material investment	1	1



(a) Time investment Test

(b) Material investment

Figure B5: Scree Plot of Eigenvalues of PCA

Both factor selection methods indicate we should use a one-factor measurement model for time and material investment. We next proceed with estimating factor loadings which are reported in Table B3. We find that the first three investment measures load positively on the latent time investment factor. The last investment measure, number of hours per day toddler spends watching tv, loads negatively on the latent time investment factor. The signs of the estimated factor loadings give us confidence that we are indeed measuring positive parenting time investment practises. The cost measures of books, toys, clothes and school all load positively on the latent material investment factor. All measures load relatively strongly on the latent factor loadings hence we retain all measures and estimate means and factor loadings using maximum likelihood. We next predict factor scores for both parental investment dimensions and further standardize the factors by the distribution of the control group.

Table B3: Estimated factor loadings time and material parental investment

	Factor Loading
Time investment	
Whether the family uses toys to play with toddler	0.577
Whether the family reads to toddler	0.549
Whether the family sings to toddler	0.564
Number of hours per day toddler spends watching tv	-0.216
Material investment	
Cost of children's books last year	0.527
Cost of children's toys last year	0.572
Cost of children's clothes last year	0.572
Cost of children's school last year	0.262

B.4. Preschool Quality

We collect data on preschool and teacher characteristics, process and structural quality. It is a priori unclear which measures best predict (perceived) preschool quality in rural China hence, in a first step, we use EFA to explore along which dimensions preschools can be best classified. Horns parallel analysis (Horn, 1965) indicates there are 5 main latent dimensions in the preschool data we collected. However, Cattell's scree plot Cattell (1966) shows that a large part of the variation in preschools can be summarized by one latent factor (Table B4 and Figure

B6). Estimated factor loadings for the 5-factor measurement model can be found in Table B5.

Table B4: EFA Factor Selection Preschool Quality

	Cattell's scree plot	Horn's parallel analysis
Preschool quality	1	5

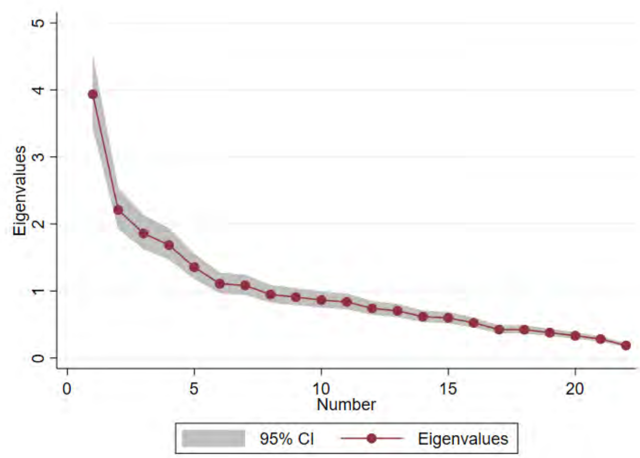


Figure B6: Preschool Quality

The pattern of estimated factor loadings in Table B5 suggest that the first latent factor is indeed a good measure of general preschool quality as it captures preschool and teacher characteristics as well as structural and process quality. Higher factor scores are associated with schools that are bigger in size, more likely to be located in township or counties as compared to villages and have younger and more educated teachers that are more likely to have received a teacher training in the past year. Higher factor scores are also associated with larger indoor and outdoor space, the availability of dormitories and breakfast and several measures of process quality such as organising exercise and science activities and reading books in class.

Variation in the second latent factor is driven by preschools with smaller pupil-teacher ratios and older teachers that receive higher salaries but score low on process quality measures. Hence

Table B5: Estimated Factor Loadings Preschool Quality

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Preschool & Teacher Characteristics					
Number of pupils	0.774	0.250	-0.103	-0.075	0.057
Share pupils receiving government need-based aid	-0.252	-0.162	0.138	0.104	0.230
Tuition fee per semester (Yuan)	0.272	0.076	-0.262	0.042	-0.166
Township preschool	0.627	0.224	-0.026	-0.060	0.138
Teacher age	-0.408	0.501	-0.136	0.300	0.012
Teacher male	-0.346	0.543	0.188	-0.088	-0.191
Teacher experience	-0.273	0.326	-0.277	0.452	0.051
Teacher monthly salary (Yuan)	-0.184	0.650	0.414	-0.119	0.028
Share of teachers with bachelor degree	0.335	0.082	0.321	-0.206	0.220
Teacher training in past year	0.519	0.130	0.273	0.013	-0.083
Structural Quality					
Pupil-teacher ratio	0.037	-0.314	-0.046	0.029	0.079
Number of activity rooms	0.715	0.281	0.058	-0.079	0.144
Outdoor play-area	0.347	0.149	-0.097	-0.043	0.061
Preschool has play room	0.245	0.118	-0.135	0.135	0.359
Preschool has exercise room	0.243	0.207	-0.434	0.185	0.239
Preschool has dormitories	0.525	0.067	-0.248	0.069	-0.287
Preschool provides breakfast	0.443	-0.064	-0.185	-0.050	-0.345
Process Quality					
Teacher reads books in class	0.331	-0.233	0.025	0.126	0.097
Teacher organizes exercise activities	0.541	-0.132	0.205	-0.027	-0.053
Teacher organizes art& music activities	0.276	-0.316	0.212	0.328	0.057
Teacher organizes science activities	0.313	0.131	0.132	0.291	-0.241
Teacher teaches social skills	0.235	-0.096	0.354	0.556	-0.058
Teacher teaches language skills	0.019	0.032	0.300	0.291	-0.022

this second latent factor might capture more informal village nurseries with small numbers of enrolled children. Variation in the subsequent latent factors is driven by a small number of items, none of which present a clear pattern. We hence proceed with a one-factor model and use estimated means and factor loadings to predict a latent preschool quality score for each preschool in the sample. Factor scores are further standardized by the distribution of the control group.

Appendix C: Additional Information on Preschool Choices and Preschool Enrolment of the Sample

We collect data on distance and travel time between each pair of program villages and preschools within each county through Gaode Map API. Table C1 shows descriptive statistics of the within county preschool choice sets and actual enrolment.

Panel A of Table C1 shows descriptive statistics of the within county preschool choice set. We find that the average minimum distance from program villages to any available preschool within the same county is about 5 kilometres which is equivalent to around 10 minutes of driving. However, a majority of households does not own a car in these rural villages, and the popular modes of transportation, either by public transportation or motorcycle, usually take much longer. The average minimum distance from program villages to any available preschools in townships or county seats within the county is about 12 km or 23 minutes driving. Hence, the difference in cost in terms of time and distance travelled between enrolling children in the closest village preschool compared to the closest township or county preschool is considerable.

Next, we look at the distance and duration of the within county preschool enrolment flows of the sample children (Panel B in Table C1). When we compare the actual enrolment choices of parents to the statistics of the preschool choice set, it becomes clear that many parents do not enrol their children in the nearest preschool but travel considerably longer to more distant preschools. The average distance between program villages and the preschools that children enrol in is 12 kilometres, or equivalent to about 21 min driving. Overall, these findings provide some suggestive evidence that for rural parents the preschool enrolment decision might be an important investment channel for which they are willing to incur extra costs.

Table C1: Within County Preschool Choice Set and Actual Enrolment

Panel A: Rural Preschool Choice Set	Mean [SD]
Minimum distance to any preschool (km)	4.785
	[7.094]
Minimum duration to a any preschool (min)	10.018
	[13.674]
Minimum distance to preschool in township or county seat (km)	12.408
	[10.598]
Minimum duration to preschool in township or county seat (min)	23.043
	[18.240]
Panel B: Actual Preschool Enrolment	
Distance to enrolled preschools (km)	11.789
	[15.962]
Duration to enrolled preschool (min)	20.853
	[25.319]

Note: Standard deviations are in the brackets.

(back to Section 4.3)

School of Economics and Finance

This working paper has been produced by the School of Economics and Finance at Queen Mary University of London

School of Economics and Finance Queen Mary University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 7356
Fax: +44 (0)20 8983 3580
Web: www.econ.qmul.ac.uk/research/workingpapers/