From Quantity to Quality: Delivering a Home-based Parenting Intervention through China's Family Planning Cadres

Sean Sylvia, Nele Warrinnier, Renfu Luo, Ai Yue,
Orazio Attanasio, Alexis Medina, and Scott Rozelle*

June 16, 2020

Abstract

A key challenge in developing countries interested in providing early childhood development programs at scale is whether these programs can be effectively delivered through existing public service infrastructures. We present the results of a randomized experiment evaluating the effects of a home-based parenting program delivered by cadres in China's Family Planning Commission (FPC) - the former enforcers of the one-child policy. We find that the program significantly increased infant skill development after six months and that increased investments by caregivers alongside improvements in parenting skills were a major mechanism through which this occurred. Children who lagged behind in their cognitive development and received little parental investment at the onset of the intervention benefited most from the program. Household participation in the program was associated with the degree to which participants had a favorable view of the FPC, which also increased due to the program.

JEL Classification: J13, I21, I28, H11

Keywords: Early Childhood Development, Parenting, China, Poverty, Family Planning

*Sylvia: University of North Carolina at Chapel Hill (email: sean.sylvia@unc.edu); Warrinnier: Queen Mary University and LICOS, KULeuven (email: n.warrinnier@qmul.ac.uk); Luo: Peking University (email: luorf.ccap@pku.edu.cn); Yue (Corresponding Author): Shaanxi Normal Univeristy (email: yueai@163.com); 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); Medina: Stanford University (email: amedina5@stanford.edu); Rozelle: Stanford University (email: rozelle@stanford.edu). The authors are supported by the 111 Project, grant number B16031. Orazio Attanasio also acknowledges support from the European Research Council (Advanced Grant AdG 695300, "Human Capital Accumulation in Developing Countries: Mechanisms, Constraints and Policies"). We thank Cai Jianhua and the China National Health and Family Planning Commission for their support on this project. We are grateful to the International Initiative for Impact Evaluation (3ie), the UBS Optimus Foundation, the China Medical Board, the Bank of East Asia, the Huaqiao Foundation, and Noblesse for project funding and to Jo Swinnen and LICOS for supporting Nele Warrinnier. We would also like to thank Jim Heckman for his support and conversations and acknowledge the support of Shasha Jumbe and the Gates Foundation's Healthy Birth, Growth and Development Knowledge Integration (HBGDki), China Program.

1 Introduction

A growing body of cross-disciplinary research highlights the importance of a child's environment in the first years of life for skill development and outcomes over the life course (Knudsen et al., 2006). This period is thought to be important for human capital accumulation both because very young children are sensitive to their environment and because deprivation during this period can have long-term consequences. Research in cognitive science suggests that malleability of cognitive ability is highest in infancy and decreases over time (Nelson and Sheridan, 2011). Due to the hierarchical nature of brain development – whereby higher level functions depend and build on lower level ones – cognitive deficiencies in early life can permanently hinder skill development. The nature of cognitive development may further lead to important dynamic complementarities in the production of human capital where early skills increase the productivity of later human capital investments and encourage more investment as a result (Cunha et al., 2010; Attanasio et al., 2015).

These mechanisms may explain findings of large long-run effects of early childhood interventions (Cunha and Heckman, 2007). Long-term follow-up studies of early childhood interventions to improve nutrition and create stimulating environments have found large and wide-ranging effects into adulthood. These studies found programs to have increased college attendance, employment, and earnings as well as cause reductions in teen pregnancy and criminal activity (Heckman et al., 2010; Walker et al., 2011; Gertler et al., 2014).

Findings from this body of research provide strong support for investments in early childhood programs (Carneiro and Heckman, 2003). Particularly in low and middle-income countries, the social returns to early intervention could be substantial due to the large number of children that are at risk of becoming developmentally delayed. Estimates indicate that 250 million children (43%) younger than 5 years old living in low-income and middle-income countries are at risk of not reaching their full development potential (Lu et al., 2016). While there are several reasons that so many children are at risk in developing countries, a significant factor is that children often lack a sufficiently stimulating environment (Black et al., 2017). Partly as a result of this evidence, Early Childhood Development (ECD) has been the subject of substantial policy advocacy, as evidenced by its inclusion in the United Nation's Sustainable Development Goals (Nations, 2015).

A key practical challenge facing policy makers, however, is how to deliver ECD programs cost effectively at scale (Berlinski et al., 2016; Richter et al., 2017). Providing ECD interventions at scale is challenging largely due to the infrastructure required to deliver services effectively to families in need, many of whom live in hard-to-reach communities such as urban slums and sparsely populated rural areas. Because building a new infrastructure to support ECD services alone would be costly, some have suggested integrating ECD

programs into existing public service infrastructures (Richter et al., 2017). For example, international agencies including the World Bank, the Inter-American Development Bank, the United Nations and the World Health Organization have called for ECD to be integrated into health and nutrition programs (Chan, 2013; Black and Dewey, 2014). Whether such a strategy can be successful is an open question. It is unclear, for example, if existing personnel who have been working in other areas and have little or no background in early childhood education can be trained to effectively deliver an ECD program. Moreover, it is often the case that public sector agencies resist new tasks, particularly if they are perceived as misaligned with the organization's existing mission (Wilson, 1989; Dixit, 2002).

We study the promotion of ECD in rural China through a home-based parent training intervention implemented by one of the world's largest bureaucracies, the China Family Planning Commission (FPC). In recent years, the Chinese government has relaxed its family planning laws and, since January 2016, has allowed all parents to conceive two children without penalty. Relaxation of the One Child Policy (OCP) and changing fertility preferences have greatly diminished the need for enforcement, and the FPC has begun to shift focus to other areas including ECD (Wu et al., 2012). Delivering ECD policies through the infrastructure of the FPC has promise but also potentially significant challenges. It is therefore unclear – even if an intervention itself is efficacious – whether it can be effectively delivered through the apparatus of the FPC.[1] This study investigates whether it is possible to re-train cadres formerly responsible for enforcing the OCP into effective parenting teachers. In other words, can the local knowledge and infrastructure of the FPC – which has been responsible for managing the quantity of human capital – be used to effectively raise the quality of human capital in China?

To study the effects of an FPC-delivered home-based parenting intervention, we conducted a cluster-randomized controlled trial across 131 villages in Shaanxi Province, located in northwestern China. We worked with the FPC to re-train 70 cadres (local officials) to deliver a structured curriculum aimed at improving parenting practices in early childhood through weekly home visits. Loosely modeled on the Jamaican Early Childhood Development Intervention (Grantham-McGregor et al., 1991), the curriculum was designed with ECD experts in China and aimed to train and encourage caregivers to engage in stimulating activities with their children.

We find that the intervention substantially increased the development of cognitive skills in children assigned to receive weekly home visits. Effects on infant skill development were accompanied by increases in both parental investment and parenting skills. Using the Generalized Random Forest method of Athey et al. (2019) to identify important sources of impact heterogeneity, we find that children who lagged behind in

[1] See China Central Television (CCTV) News report: How will a Million Family Planning Workers Transition? https://youtu.be/84WIe1C3XTM

their cognitive development and received little parental investment at the onset of the intervention benefited most from the program. Although the average effect of the program was diminished by imperfect compliance, we find evidence that one of the primary factors hindering compliance – unfavorable public perception of the FPC – was also significantly reduced as a result of the program. This suggests that compliance may improve over time if implemented by the FPC.

Our findings add to an emerging literature studying how ECD can be integrated into existing infrastructure in developing countries to facilitate delivery at scale. Attanasio et al. (2014) found that a parenting intervention integrated into an existing conditional cash transfer program in Colombia and delivered by local volunteers successfully improved cognitive development outcomes, and, like the program we study in China, did so primarily through increased parental investments (Attanasio et al., 2015). Again in Colombia, Attanasio et al. (2018) analyze the impact of a stimulation intervention implemented within an existing programme promoted by the Colombian government and show that it has a sizable impact on children developmental outcomes. In Pakistan, Yousafzai et al. (2014) find significant improvements in early childhood outcomes of children enrolled in a parenting intervention integrated in a community-based health service and find that effects persist 2 years after termination of the parenting intervention (Yousafzai et al., 2016). Our study adds to the literature by providing evidence on the effectiveness of an ECD intervention integrated into local government services in China: specifically whether the infrastructure and personnel of the FPC can effectively implement a home-based parenting program and reduce the high prevalence of cognitive delay among infants and toddlers in rural China.

The remainder of the paper is structured as follows. In the next section we discuss the FPC and how their role is changing with the abolishment of the One Child Policy. In section 3 we describe the experimental design and data collection. In section 4 we report findings of the impact evaluation of the parenting intervention. Section 5 concludes.

2 Background: The Changing Role of the FPC

The Family Planning Commission (FPC)[2] is the entity responsible for the implementation of population and family planning policies in China. From 1980, a large part of the agency's mandate included enforcement of the One Child Policy – a policy comprised of a set of regulations governing family size.[3] Although there were several, now well-documented, unintended consequences of the policy, the government at the

[2] In March 2013, the National Population and Family Planning Commission was merged with the Ministry of Health to form the current National Health and Family Planning Commission. Since March 2018, the ministry is called the National Health Commission.

[3] Despite its name, most families were not restricted to having only one child. In many rural areas, families were allowed two children and there were a number of other exemptions including for minority groups and for parents who worked in high-risk occupations. See Hesketh et al. (2005) and Hesketh et al. (2015) for good overviews of the policy and implementation.

time considered population containment necessary to improve living standards as the country faced an impending baby boom (Hesketh et al., 2005).

The implementation of China's One China Policy required close interaction between families and local FPC cadres to ensure universal access to contraceptive methods, to monitor for violations, and to enforce penalties. Although details of how the policy was implemented varied across regions and time, at its most intense phase of implementation families were required to obtain birth permits before pregnancy and births were to be registered with the local FPC cadre. Once families met their number of allowed children, FPC officers often encouraged or forced sterilization (Greenhalgh, 1986). If women became pregnant without a birth permit, FPC facilities were used for abortions (both voluntary and not). The FPC also enforced penalties for out-of-plan births which included substantial fines and loss of employment.

Given the numerous and complicated set of policy instruments, and the close interaction with families that this entailed, implementation of the One Child Policy required a large bureaucracy. As of 2005, the FPC had more than 500,000 administrative staff and more than 1.2 million village-level FPC operatives.[4] In 2016, the budget supporting the FPC's activities exceeded 8.85 billion dollars.[5] However, after debates in recent years about the necessity of the One Child Policy's continuation, the government announced in October 2015 that the policy would be formally terminated as of January 1, 2016.[6] Termination of the policy also has called into question the future role of the FPC.[7]

Some have argued that an appropriate future focus of the FPC would include early childhood care and education, which falls within the technical purview of the agency (Wu et al., 2012). Currently, responsibility for providing these services is spread across multiple entities, which, in practice, has led to a gap in service provision (Wu et al., 2012). Whether the FPC would be able to effectively fill this role is an open question, however. On one hand, the FPC has the ideal infrastructure to provide early childhood services: a large, well-functioning organisation with representation in every village and community in the country; a relatively well-educated work force; and the ability to maintain information on every family and child. On the other hand, it may be difficult for FPC cadres to retrain and effectively deliver ECD services. More significantly, the agency's history and reputation could limit its effectiveness. Although the enforcement of the policy relaxed over time, the agency's at times draconian measures may have created lasting social animosity toward the family planning commission that could hinder its effective delivery of ECD services. Moreover, given that

[4]See NPFPC, 2006, Statistical Bulletin of Forth National Population and Family Planning System Statistical, http://www.nhfpc.gov.cn/guihuaxxs/s10741/201502/f68e73331a9147e78209ab81bd156a39.shtml

[5]Includes funding for health and family planning activities. See NHFPC, 2016, The Departmental budget report of National Health and Family Planning commission of the PRC, http://www.nhfpc.gov.cn/caiwusi/s3574/201604/3582098e060144148a1e3b4f3f1a4fe0.shtml

[6]The Central Committee of the Communist Party of China, 2015. Bulletin of Fifth Plenary Session of 18th CPC Central Committee.

[7]See Sonmez, F., Wall Street Journal, 2015. After the One-Child Policy: What Happens to China's Family-Planning Bureaucracy? http://blogs.wsj.com/chinarealtime/2015/11/12/after-the-one-child-policy-what-happens-to-chinas-family-planning-bureaucracy/

the agency is responsible for other tasks, it is unclear if FPC cadres would allocate (or be directed to allocate) sufficient effort to the parenting program to make it effective.

3 Experimental Design and Data Collection

3.1 Sampling and Randomization

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.

Administration in China's rural areas is organized in a three-tier system comprised of villages (lowest tier), townships (middle tier), and county (upper tier).The average population of villages in our sample region is around 1,600. There are approximately 12 villages within each township and 10 townships per county. To identify the sample, we first 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, government data were used to compile a list of all villages reporting a population of at least 800 people. We then randomly selected two villages from the list in each township. These exclusion criteria were chosen to ensure a rural sample and increase the likelihood that sampled villages had a sufficient number of children in the target age range. Our final sample consisted of 131 villages total[8]. All children in sample villages between 18 and 30 months of age were enrolled in the study. At baseline, a total of 592 children were sampled.

Following baseline data collection (described below), 65 villages were randomly assigned to the parenting intervention group and the remaining 66 to a control group. The randomization procedure was stratified by county, child cohort, and experimental group of an earlier trial. Each trainer was assigned a maximum of four families chosen randomly from rosters in treatment villages to be enrolled in the program. In treatment villages, a total of212 children were enrolled and the remaining 79 were not. Because these children were randomly selected, the two groups have the same characteristics in expectation. In the analysis, we test for spillover effects on these children in treatment villages who were not selected to participate.

3.2 Parenting Program

Parenting trainers, selected by the FPC from among their cadres in each township, delivered a structured curriculum through weekly home visits to households in treatment villages for a period of six months

[8]One of the villages had no children in the target age range and was therefore dropped prior to randomization

(from November 2014 to April 2015). Based loosely on the Jamaican home visiting model (Grantham-McGregor et al., 1991) and adapted by child development psychologists in China to the local setting, the goal of the intervention was to train caregivers to interact with their children through stimulating and developmentally-appropriate activities.

The curriculum delivered by the parenting trainers was developed by the research team in collaboration with the FPC and outside ECD experts in China. The curriculum was stage-based and fully scripted. Weekly age-appropriate sessions were developed targeting children from 18 months of age to 36 months of age. Each weekly session contained modules focused on two of four total developmental areas: cognition, language, socio-emotional, and (fine and gross) motor skills. Every two weeks, caregivers would encounter one activity from each category. In addition to developmental activities, the curriculum also included one weekly module on child health/nutrition.

During sessions, parent trainers were trained to introduce caregivers to the activity and assist caregivers to engage in the activity with their child. Typically the only caregiver that participated was the primary caregiver (usually mother or grandmother), though other caregivers sometimes observed. 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.

Parenting trainers were selected and deployed by the FPC office in each township. Summary statistics on trainer characteristics are shown in Appendix Table A1. Around 60 percent of the parenting trainers deployed by the FPC office were men. The majority of parenting trainers were married and had children themselves. The parenting trainers were well educated with most of them having enjoyed a community college higher education and around 29 percent had obtained a bachelor degree. On average, parenting trainers were 34 years old and had worked 12 years for the Family Planning Commission. FPC offices assigned parent trainers to enrolled families in their township. Most trainers were assigned families in only one village.

Fully scripting the curriculum eliminated the need for extensive training of parent trainers. All parenting trainers underwent an initial, centralized 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.

3.3 Data Collection

We conducted our baseline survey in October 2014 and our follow-up survey in May 2015. Teams of enumerators collected detailed information on children, caregivers and households. Each child's primary caregiver was identified and administered a survey on child, parent and household characteristics including each child's gender, birth order, maternal age and education. Each child's age was obtained from his or her birth certificate. The primary caregiver was identified by each family as the individual most responsible for the infant's care (typically the child's mother or grandmother). In both the baseline and endline surveys, we collected data on children's cognitive and psychomotor development; children's social-emotional behaviour; and parenting skills and investments. Detailed data on compliance (household visits completed) was also collected throughout and after the intervention.

Cognitive and Psychomotor Development Children's cognitive, psychomotor and social-emotional development were assessed in each round. At baseline, all children were assessed using the Bayley Scales of Infant Development (BSID) Version I, a standardised test of infant cognitive and motor development (Bayley, 1969). The test was formally adapted to the Chinese language and environment in 1992 and scaled according to an urban Chinese sample (Yi et al., 1993; Huang et al., 1993). Following other published studies that use the BSID to assess infant development in China (Li et al., 2009; Chang et al., 2013; Wu et al., 2011), it was this officially adapted version of the test that was used in this study (Yi, 1995). All BSID enumerators attended a week-long training course on how to administer the BSID, including a 2.5 day experiential learning program in the field. The test was administered in the household using a standardised set of toys and detailed scoring sheet. The BSID takes into consideration each child's age in days, as well as whether he or she was premature at birth. These two factors, combined with each child's performance on a series of tasks using the standardised toy kit, are used to construct two sub-indices: the Mental Development Index (MDI), which evaluates memory, habitation, problem solving, early number concepts, generalisation, classification, vocalisation and language to produce a measure of cognitive development; and the Psychomotor Development Index (PDI), which evaluates gross motor skills (rolling, crawling and creeping, sitting and standing, walking, running and jumping) and fine motor skills to produce a measure of psychomotor development (Bayley, 1969).

Because the BSID-I is not designed to assess outcomes for children older than 30 months, only children aged 30 months or under at follow-up (approximately half of the sample) were administered the BSID in the follow-up survey. Older children were assessed using the Griffith Mental Development Scales (GMDS-ER 2-8) (Luiz et al., 2006), which has been shown to be comparable in its assessment of early childhood

development to the BSID-I (Cirelli et al., 2015).[9]

Enumerators were trained on how to administer the Griffith Mental Development Scales. As with the BSID, a standard activity kit is used to test different skill sets of children and enumerators score children on a standardised form based on their performance on tested activities. The GMDS-ER 2-8 comprises six sub scales: locomotor, personal-social, language (receptive and expressive), hand and eye coordination, performance, practical reasoning.[10]

For the analysis, raw scores are standardised separately by sub-index. Since raw scores are increasing in age, we compute age-adjusted z-scores using age-conditional means and standard deviations estimated by non-parametric regression. This non-parametric standardisation method is less sensitive to outliers and small sample size within age-category and yields normally distributed standardised scores with mean zero across the age range (in months)(Attanasio et al., 2015).[11]

Socio-emotional Behaviour In each wave we also assessed children's social-emotional behaviour using the Ages and Stages Questionnaire: Social Emotional (ASQ:SE) (Squires et al., 2003). The items in this questionnaire (which vary by age) measure a child's tendency towards a set of behaviors such as ability to calm down, accept directions, demonstrate feelings for others (empathy), communicate feelings, initiate social responses to parents and others, and respond without guidance (move to independence). Main caregivers were asked to indicate whether the child exhibits these behaviors *most of the time*, *sometimes*, or *never*. Depending on the desirability of the behavior, answered are scored either 0, 5, or 10 points. Children who score 60 or more are considered to require further assessment for social-emotional problems.

Parenting Skills and Investment The parenting curriculum was designed to affect child development by increasing parenting skills and investment of caregivers in the development of their children. We measured parenting skills at baseline and follow up by asking the primary caregiver a series of questions on parenting knowledge and confidence. These included questions about the importance of different activities such as reading and playing with their children and caregiver confidence in engaging in these activities. Caregivers responded to these questions using a 7-point likert scale. Parental investment was measured by asking whether the main caregiver engaged in a set of child-rearing activities, such as story-telling and playing with toys, the previous day and how many children's books they have in the house.

Compliance Information on compliance – including whether the weekly parenting sessions took place and, if not, the reason they did not take place – as well as details of the interaction were collected on a monthly basis from caregivers and on a weekly basis from parenting trainers through telephone interviews.

[9] The Pearson correlation coefficient between the BSID and GMDS is found to be higher than 0.8.

[10] The last sub-scale of the GMDS-ER, practical reasoning, is only used to assess development of older children, hence was not registered to this particular age group. Furthermore, in the analysis we ommit the GMDS-ER language subscale as receptive and expressive language skills are not explicitly tested by the BSID I and we want to have comparable measures across the two age cohorts.

[11] The non-parametric method is described further in the Web Appendix B.4. of Attanasio et al. (2015).

In our analysis, we use parenting trainer reports as these data are more complete. The difference in average compliance for these two measures is insignificant and the two measures are highly correlated (correlation of 0.69).

3.4 Baseline Characteristics, Balance, Attrition

Summary statistics and tests for balance across control and treatment groups 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 the study arms are balanced.[12] Appendix Table A2 shows that characteristics of untreated children in treatment villages (the "spillover group") are also balanced with those of children in the treatment and control groups.

Children in our sample are on average just over 24 months old at the start of the program. Less than 5% of children are born with low birth weight. A large part of the children in our sample are first born 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.[13] 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.

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 as appropriate) have at least 9 years of formal schooling. On average households report being somewhat indifferent in their feelings toward the FPC at baseline.[14]

Baseline statistics on parental inputs shown in Panel C of Table 1 show 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 or singing to their child the previous day.

[12] 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 0.564.

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

[14] We asked caregivers to rate their perception of local Family Planning Commission on a 5-point scale (1 *very much like*; 2 *like*; 3 *Neither like nor dislike*; 4 *dislike*; 5 *very much dislike*).

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.

4 Estimation of Program Effects

Given 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 randomization procedure) we condition our estimates on randomization 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:

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

where Y_{ijt} is an outcome measure for child 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. To estimate spillover effects we use the same specification but replace treated children with untreated children in treatment villages in the estimation sample. Because we estimate treatment effects on multiple outcomes, we present p-values adjusted for multiple hypotheses using the step-down procedure of Romano and Wolf (2005, 2016) which controls for the familywise error rate (FWER)[15].

We estimate program effects both separately by age cohort and on the full sample pooling both cohorts together. Because different assessments were used for the cohorts at endline, we construct a combined index of infant skill development that allows us to estimate effects on the full sample. To construct this index, we follow Heckman et al. (2013) and Attanasio et al. (2015) and develop a dedicated measurement system relating the observed infant development outcome measures in both cohorts to a latent infant skill factor. We assume that the measurement system is invariant to treatment assignment which implies that any observed treatment effect on measured development outcomes results from a change in the latent skill and not from a

[15] To compute adjusted p-values, we follow the algorithm described in Romano and Wolf (2016) using the RWOLF command in Stata (Clarke, 2018). In estimating treatment impacts on infant skills, p-values are adjusted across all 8 outcomes for the two cohorts. For effects on secondary outcomes, parental investment and skills, p-values are adjusted within each group corresponding to investments and skills separately following the conceptual framework in Section 5.2.

change in the measurement system.[16] Hence, for each cohort we estimate following dedicated measurement system at baseline and follow-up:

$$y_{im}^{\theta} = \mu_m^{\theta} + \theta_i' \lambda_m^{\theta} + \delta_{im}^{\theta} \tag{2}$$

with y_{im}^{θ} the observed m^{th} measure for child i; μ_m^{θ} the mean of the m$^{\text{th}}$ measure and λ_m^{θ} the loadings of the factor for measure m. The measurement error δ_{im}^{θ} is the remaining proportion of the variance of the outcome measures m that is not explained by the factor and is assumed to be independent of the latent infant skill factor θ and to have a zero mean.[17]

After estimating the measurement system for each cohort separately we use the estimated means and factor loadings to predict a factor score for each child i in the sample using the Bartlett scoring method (Bartlett, 1937)[18]. The predicted infant skill factors are standardized non-parametrically for each age-month group by cohort and we control for cohort fixed effects in our pooled regression specification.

In the same spirit as the creation of a latent infant skill factor, we estimate a dedicated measurement system relating all observed measures of parental investment behaviour and parenting skills to latent factors. We estimate following system of equations for baseline and follow-up:

$$y_{im}^{P} = \mu_m^{P} + P_i' \lambda_m^{P} + \delta_{im}^{P} \tag{3}$$

$$y_{im}^{I} = \mu_m^{I} + I_i' \lambda_m^{I} + \delta_{im}^{I} \tag{4}$$

with y_{im}^{P} and y_{im}^{I} the observed m^{th} measure of parenting skill or parental investment of child i; μ_m^{P} and μ_m^{I} the mean of the m$^{\text{th}}$ measure and λ_m^{P} and λ_m^{I} the loadings of the factor for measure m. To implement the dedicated measurement system described above we first perform an exploratory factor analysis (EFA), reported in Appendix B, in order to identify in a preliminary step the relevant measures and their allocation to the latent factor as shown in Table B1 - Table B4. The measurement system for the latent parenting skill

[16] More formally, this assumption implies that the measurement system intercept, factor loadings and distribution of measurement errors are the same for the control and the treatment group

[17] **Table B5** in the appendix shows the measurement system for the latent infant skill factor at baseline and follow-up. The first column in this table reports factor loadings. We normalized the factor loading of the first measure in both periods and cohorts to one. Hence, at baseline, the scale of the latent infant skill factor is determined by the Bayley Mental Development Index. At follow up, the scale of the latent infant skill factor is determined by the Bayley Mental Development Index for the younger cohort, and by the Griffith Performance scale for the older age cohort. The second column of the table shows estimates for how much of the variance is driven by signal relative to noise. The signal-to-noise ratios for the m^{th} measure of child development is calculated as:

$$S_m^{\theta} = \frac{\lambda_m^2 Var(\theta)}{\lambda_m^2 Var(\theta) + Var(\delta_m)}$$

These calculations show that Bayley and Griffith measures derived form objective testing by trained enumerators have relatively high signal-to-noise ratios while the signal of the *ASQ: Social-Emotional*, a measure based on caregiver response, is relatively poor.

[18] Bartlett's scoring method is based on GLS estimation with measures as dependent variables and factor loadings as regressors.

factor and parental investment factor at baseline and follow-up can be found in Appendix B Table B5. The predicted parenting skill factor and parental investment factor are standardized by the distribution of the control group.

5 Impact of the Parenting Intervention

5.1 Average Treatment Effects on Infant Skills

Pooling the two cohorts, Figure 1 plots the kernel density estimates of the latent infant skill distribution at baseline and follow-up by treatment assignment. At baseline, the infant skill distribution of infants in treatment and control villages overlap and a Kolmogorov-Smirnov (K-S) test indicates that the two distributions are similar (p-value = 0.828). At follow-up, the infant skill distribution is shifted to the right in the treatment group. A K-S test rejects the equality of distributions in the treatment and control groups with a p-value of 0.029.

Table 2 presents the average treatment effects on infant skills. Pooling cohorts, we estimate that the parenting program led to an overall average increase of 0.246 standard deviations in infant skill (bottom row). Estimating effects separately by cohort, we find that the parenting intervention significantly increased cognitive skills as measured by the Mental Development Index of the Bayley assessment scale for the younger age-cohort and by the Griffith assessment scales of Performance and Personal-Social for the older age-cohort. The 6-month intervention led to a significant increase of 0.292 standard deviations in cognitive development in the younger cohort and an increase of 0.280 standard deviations for the older cohort. We find no significant program effects on child psychomotor development or on social-emotional outcomes. These results are similar to the finding of Attanasio et al. (2014), who report that their home-based parenting intervention in Colombia led to an increase of 0.26 standard deviations in cognitive development but no significant improvement in psychomotor development. Despite similar effect sizes of both programs, the Colombia study lasted one year longer (18 months in total) and enrolled younger children (12-24 months).

5.2 Mechanism: Effect on Parenting Skills and Parental Investment

To motivate the mechanisms through which the parenting intervention may have affected infant skills, consider the following general production function of early skill formation:

$$\theta_{t+1} = f_{t+1}(\theta_t, I^T_{t+1}, I^P_{t+1}, P_{t+1}, X_t). \tag{5}$$

Here, θ_t and θ_{t+1} are vectors of infant skills at baseline and follow-up respectively, I^T_{t+1} are direct investments from the treatment (i.e. time spent with the child during weekly visits), I^P_{t+1} are parental investments during the intervention period, P_{t+1} are parenting skills during the intervention period, and X_t a vector of household characteristics.

This production function illustrates several mechanisms through which the intervention may have affected infant skill. First, the intervention could have a *direct* impact on infant skill formation through the weekly interactions with the parenting trainers (investment from the treatment itself, a shift in I^T_{t+1}). Alternatively, the intervention may have *indirect* effects by affecting either (a) parental investment (I^P_{t+1}) or (b) the effectiveness of parental investment through an increase in parenting skills (P_{t+1}). Although the intervention was designed to improve the quantity and quality of infant-caregiver interactions it is not a priori clear that parents would spend more time with their children. Parental investment could be crowded-out as a result of the intervention if parents see the intervention as an in-kind transfer and hence re-optimize the allocation of the household resources.[19]

Our data allow us to estimate the causal effect of the intervention on two of these four mechanisms: parental investments and on parenting skills. Assuming measurement error is sufficiently small, no treatment effects on parental investment would suggest that the main mechanism for program effects is through a direct effect of the program. Effects on these two indicators, however, would not rule these out as potential channels of impact.

Kernel density estimates of the latent parental investment factor and the latent parenting skill factor at baseline and follow-up are plotted in Figure 2 by treatment assignment. At baseline both the parental investment factor and parenting skill factor have a similar distribution for control and treatment villages (confirmed by K-S test p-values of 0.973 and 0.889 respectively). At follow-up we find that the distribution of the parental investment factor in the treatment villages has drastically shifted to the right. This visual evidence is also supported by a strong K-S test rejection of the equality of the two parenting investment factor distributions with a p-value < 0.001. We see a more moderate shift in the distribution of the parenting skill factor. Nevertheless, the distributional shift is significant (p-value=0.003) and we find again that caregivers in treatment villages have improved parenting skills along the entire ability distribution.

Average treatment effects on the secondary outcomes can be found in Table 3. We find that the program significantly increase parenting skills with an overall increase of 0.323 standard deviation in parenting skill found in treatment villages (Panel A). In terms of individual components, caregivers in treatment

[19] An additional potential mechanisms is that the intervention could change the production technology by shifting the productivity parameter. Attanasio et al. (2014) use data from an intervention in Colombia to explicitly test for this mechanism and do not find evidence for this channel. Following this result, we do not test for this mechanism here (as we focus on reduced-form results), but assume that this channel is negligible in our interpretation of mechanisms.

households report a stronger belief in the importance of reading for child development and more confidence in their ability to read to their children. We also find some evidence that parents in treatment villages are more confident (less nervous) about their ability to care for their children[20]. The intervention had no effect on parental beliefs about the importance of play for child development nor on parental beliefs about their communication skills with their offspring.

We also find large effects on parental investment with overall parental investment increasing with 0.825 standard deviations in treatment villages (Panel B). The parenting intervention increased the time caregivers spend with their children actively engaging in age-appropriate developmental activities such has reading and singing. Furthermore, we find that treatment households had significantly more children's books in their homes at the end of the program compared to the households in the control group. We find no evidence of crowding-out of parental investment as a result of the parenting intervention as children in treatment households did not significantly spend more time watching tv or playing by themselves.

Overall this evidence suggests that parents are investing considerably more effort into parenting and have gained some better parenting skills as a result of the intervention. This evidence suggests that an important mechanism contributing to the effectiveness of the intervention was a change in parenting behavior, which was the aim of the parenting intervention and is in line with findings of Attanasio et al. (2015).

5.3 Compliance and Dose-Response Estimation

On average, 16.4 visits (out of 24 total planned visits) were completed for each household during the course of the study based on reports from parent trainers. To assess the drivers of incomplete compliance, we regress the number of reported household visits on child, family, and trainer characteristics as well as the distance from the village to the closest FPC office. The estimated correlates of compliance can be found in Table 4a.

Compliance is most strongly correlated with four factors: whether the child is male, whether a child suffered cognitive delay at the start of the intervention, distance from the village to the FPC office in the township, and caregiver perception of the FPC. Male children receive on average slightly more household visits. Children who were cognitive delayed (measured as BSID< 80) received on average one to two household visits less compared to children who were at a more normal developmental stage at the start of the intervention. Households located further away from FPC offices located in township centres also tended to receive fewer household visits. This could be due to either supply-side compliance failure as parenting trainers chose to visit remote households less frequently or reflect household characteristics correlated with

[20] When controlling for the familywise error rate of the parenting skill measures using the Romano-Wolf (2005) stepdown procedure this individual component is no longer significant at conventional levels

remoteness. However, observed household characteristics are weakly correlated with distance in our sample (Table 4b) suggesting that negative correlation with distance is more likely due to supply-side shirking.

Once all variables are included in the compliance regression, the most important demand-side factor associated with compliance appears to be whether households had an unfavorable view of the FPC at baseline. Households with a more unfavorable view of the agency completed significantly fewer visits. If the program were to be implemented in the future, however, this may become less of an obstacle to implementation as we find that the program itself has a significant positive effect on public perception of the FPC as reported in Table 5. The estimated average treatment effect of the intervention on the household's reported negative perception of the FPC (on 6-point likert scale) at the end of the parenting program is -0.316 and significant at the 5% level.

Given imperfect compliance, we present estimates of the dose-response relationship between the number of completed household visits and our main outcomes of interest (infant skill, parenting skill, and parental investment). As compliance to the parenting program is a choice variable the initial randomization does not preclude selection bias on treatment intensity. In estimating the dose-response relationships we therefore need to control for potential sources of confounding variables that cause selection bias. Traditionally, in the literature, this is achieved by instrumenting compliance with treatment assignment. This, however, implicitly assumes that the dose-response function is linear in the number of household visits. We relax this assumption and allow for a concave relationship. More specifically, we use a control function method first assuming a linear relationship and than allowing for a concave relationship by adding a squared term for household visits completed. Control function methods rely on similar identification conditions to two stage least squares (2SLS) and coincide with 2SLS in a linear model.[21] Identification requires instruments that are relevant and can be excluded from the production and investment functions under reasonable assumptions. For each of the outcomes of interest, we instrument the number of household visits with the treatment assignment, the distance between the village and the FPC township office, and the interaction between these two variables. The implicit assumption here is that treatment intensity is related to distance of the household to the Family Planning Office but that the distance measure does not affect the skill accumulation process nor the parental investment decision, conditional on treatment intensity.[22] We use ordinary least-squares (OLS) to estimate the first stage equations for each of the three main outcomes:

$$V_{ijt} = \alpha_1 + \beta_1 T_{jt} + \beta_2 T_{jt} * D_{jt} + \beta_3 D_{jt} + \gamma_1 Y_{ij(t-1)} + \tau_s + \xi_{ij} \tag{6}$$

[21] We refer to Wooldridge (2015) for an overview of control function methods in applied econometrics.
[22] Linear estimates of the dose-response relationship between the number of completed household visits and cognitive development outcomes are similar when instrumenting compliance with only treatment assignment.

where V_{ijt} is the number of completed household visits for child i in village j at follow-up; T_{jt} is a dummy variable indicating the treatment assignment of village j; D_{jt} the distance of village j to the Family Planning Office; $Y_{ij(t-1)}$ 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. Estimates of the first stage regressions can be found in Table A4 in the Appendix. Next, using the estimated residuals, $\hat{\xi}_{ij}$, we proceed to estimate the second stage equations for the three main outcomes:

$$Y_{ijt} = \alpha_2 + \beta_4 V_{ijt} + \beta_5 \hat{\xi}_{ij} + \gamma_2 Y_{ij(t-1)} + \tau_s + \eta_{ij} \tag{7}$$

$$Y_{ijt} = \alpha_3 + \beta_6 V_{ijt} + \beta_7 V_{ijt}^2 + \beta_8 \hat{\xi}_{ij} + \beta_9 \hat{\xi}_{ij}^2 + \gamma_2 Y_{ij(t-1)} + \tau_s + \upsilon_{ij} \tag{8}$$

where Y_{ijt} is an outcome measure for child i in village j at follow-up; $Y_{ij(t-1)}$ is the outcome measure for child i at baseline; V_{ijt} the number of completed household visits at follow-up and V_{ijt}^2 the squared number of completed household visits at follow-up; $\hat{\xi}_{ij}$ the estimated residual of the first stage equation and $\hat{\xi}_{ij}^2$ the squared residual; τ_s is a set of strata fixed effects. We adjust standard errors for clustering at the village level using the Liang-Zeger estimator.

Table 6 shows control function estimates of the dose-response relationships. In Columns (1), (3) and (5) we assume a linear relationship between the number of completed household visits and the latent infant skill, parenting skill and parental investment factors. We estimate that each session completed increases infant skill with 0.013 standard deviations, parenting skill with 0.019 standard deviations and parental investment with 0.049 standard deviations. Results from Column (2), (4) and (6) which allow for non-linearity do not suggest that these relationships are concave. Assuming a linear relationship up to 24 household visits, these estimates suggest that under full compliance we would see infant skill increase by 0.312 standard deviations, parenting skill by 0.456 deviations and parental investment by 1.176 standard deviations.

5.4 Impact Heterogeneity

The production function of early skill formation (Equation 5) suggests that heterogeneity in treatment effects of the parenting program could arise from a large variety of sources. Treatment effects could differ across children due to differences in initial skills as well as differences in household and community characteristics that affect participation in and efficacy of household visits, or how caregivers respond to household visits.

The variety of potential sources of heterogeneity creates an empirical challenge since – as is the case for most randomized trials – increasing sample size to be sufficiently large to provide enough power to test heterogeneity across a large number of dimensions would be prohibitively costly. While the number of tests performed could be limited ex-ante, this approach would increase the likelihood that important sources of heterogeneity are missed (Almås et al., 2018).

To examine heterogeneity in a principled way, we therefore use recently developed machine learning approaches to inform our analysis of heterogeneous treatment effects. Specifically, we first use the Generalised Random Forest (GRF) method developed in Athey et al. (2019) to predict subgroups in which there is a significant amount of treatment effect heterogeneity and use these predictions as a guide in a more traditional heterogeneity analysis. This allows us to limit heterogeneity tests (and hence the probability of over-rejection) while minimising the probability that important sources of heterogeneity are neglected.

Predicting Impact Heterogeneity Using Generalised Random Forest Analysis

The first step in our analysis of heterogeneity is to assess which observable characteristics measured at baseline predict differences in treatment effects of the parenting program. Building on methods that extend regression tree and random forest algorithms from a tool for general prediction to an algorithm that can estimate conditional average treatment effects (CATE) for different sub-groups of the population (Athey and Imbens, 2016; Wager and Athey, 2018), Athey et al. (2019) introduce the Generalized Random Forest (GRF) algorithm, which produces estimates that are consistent and asymptotic normally distributed with a variance that can be estimated, making inference possible.[23] GRFs keep the typical structure of traditional Random Forests but, instead of aggregating across all trees in a forest by taking the average, estimate a weighting function and use these weights to solve local moment equations. We use the GRF algorithm to build a Causal Random Forest (CRF) to estimate conditional average treatment effects (CATE):

$$\tau(X) = E[Y(T=1) - Y(T=0)|X = x] \tag{6}$$

where Y is the outcome variable and T indicates treatment assignment which is assumed independent of unobservable variables conditional on the observable covariates, X. As our sample is relatively small and random forest methods perform better in larger samples (Davis and Heller, 2017), we use the GRF algorithm

[23] To enable statistical inference in the GRF algorithm, Athey et al. (2019) use "honest trees." Honest trees split the training data into two separate subsamples: one to perform the splits (generate the tree) and one to make predictions. Observations in the estimation data are then applied directly to the "terminal nodes" (leaves) of the tree and treatment effects are estimated by comparing treatment and control observations within each terminal node. This procedure produces estimates that are consistent and asymptotically normal.

to build a CRF[24] as a pre-regression analysis, in line with the strategy used by Carter et al. (2019).[25] We select 12 baseline characteristics for this prediction problem, listed in Table 7. After training the GRF algorithm on the selected characteristics we investigate which of these characteristics is relatively more important in predicting treatment heterogeneity.

Before analysing whether certain subgroups benefited more or less from the parenting intervention it is useful to check how much treatment heterogeneity in infant skills at program completion we observe in our sample. The distribution of predicted out-of-bag CATE's[26], shown in Figure 3, indicates substantial variation in how children responded to the home visiting intervention. The predicted treatment intensity varies between 0.07 and 0.45 of a standard deviation in infant skills. The cumulative distribution of the estimated out-of-bag CATEs (Figure 4), shows that children in the bottom quartile of the CATE distribution are estimated to have gained between 0.07 and 0.14 standard deviations in infant skill at endline while infants in the top quartile gained between 0.34 and 0.45 standard deviations. A simple approach proposed by Wager and Athey (2018) to test more formally for heterogeneity involves grouping observations according to whether their out-of-bag CATE estimates are above or below the median CATE estimate and than estimating average treatment effects in these two subgroups separately. We find that the estimated difference between the two groups is relatively large at 0.334 standard deviations of infants skill and statistically significant (p-value=0.047). The average treatment effect of 0.23 standard deviations shown in Table 2 hence hides considerable variation in treatment effects for children within in the treatment group.

To explore which specific sub-groups benefited more from the intervention at endline, we first consider the variable importance calculated by the GRF algorithm and shown in Table 7. This measure captures the percentage of *importance* each observable characteristic has in the forest in terms of the frequency with which the variable is used as a splitting variable in the forest. The higher the percentage, the better that variable is in predicting treatment heterogeneity. We find that the level of parental investment at baseline is by far the best predictor of treatment effect heterogeneity. Other predictors of heterogeneity are infant skills at baseline and the distance to the FPC office. In Figure 5 we next plot the estimated out-of-bag CATEs

[24] Borrowing notation from Wager and Athey (2018) we give a short description below of the prediction problem. The GRF algorithm makes predictions as an average of b trees as follows: (1) For each b = 1,...B, draw a subsample $S_b \subseteq \{1,..,n\}$; (2) Grow a tree via recursive partitioning on each such subsample of the data; and (3) Make predictions

$$\hat{\tau}(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{n=1}^{n}\frac{Y_i 1(\{X_i \in L_b, i \in S_b\})}{|\{i : X_i \in L_b, i \in S_b\}|} \tag{9}$$

where $L_b(x)$ denotes the leaf of the b-th tree containing the training sample x.

[25] For a technical explanation of the GRF algorithm we refer to Athey et al. (2019), for a less technical explanation and examples of the application of the GRF algorithm to policy impact evaluations we refer to Davis and Heller (2017) and Carter et al. (2019). Information about the implementation of the GRF algorithm in R can be found at https://cran.r-project.org/web/packages/grf/grf.pdf

[26] In the case of out-of-bag prediction the estimated CATE's only consider trees for which the observation is not used as part of the training set: $i \notin S_b$.

from the GRF estimation along the distribution of these three characteristics.[27] A clear pattern emerges from the first two scatter plots. Overall, higher estimated CATEs are found for infants that were more at a disadvantage at the start of the intervention. We find that higher estimated program impacts are associated with lower parental investment at baseline and lower infant skills at baseline. Distance from the household to the Family Planning Office also is an important predictor of impact heterogeneity but the scatterplot shows a less clear pattern between the estimated out-of-bag CATEs and the distance measure. Based on the results of the supervised learning algorithm we proceed in the next section with testing for heterogeneous program impacts along these three dimensions.

GRF-Informed Heterogeneity Analysis

To test whether the parenting program was more effective for infants who faced an initial relative disadvantage at the start of the intervention or lived in households further away from the Family Planning Offices, we define three new variables indicating relative disadvantage in the dimensions of initial parental investment, infant skill and distance. More precisely, we define for each of these dimensions a dummy variable indicating whether the children were below a certain threshold in the baseline distribution. We define the threshold for each dimension based on how the estimated out-of-bag CATEs from the GRF analysis vary across the baseline distribution of each variable. For both the parental baseline investment and distance measure the scatter plots of Figure 5 suggest non-linearity in the treatment heterogeneity, specifically sharp declines in estimated CATEs at lower tails of the pre-intervention distribution. We therefore define an indicator for being in the first quartile of the pre-intervention distribution. Using these new indicator variables, we estimate intention-to-treat effects of the parenting intervention using ordinary least-squares with the following ANCOVA specification:

$$Y_{ijt} = \alpha_1 + \beta_1 T_{jt} + \beta_2 T_{jt} Q_{ij(t-1)} + \beta_3 Q_{ij(t-1)} + \tau_s + \epsilon_{ij} \tag{7}$$

where Y_{ijt} is an outcome measure for child i in village j at follow-up; T_{jt} is a dummy variable indicating the treatment assignment of village j; $Q_{ij(t-1)}$ is the relevant indicator defined using the baseline characteristic of interest; $T_{jt}Q_{ij(t-1)}$ the interaction of treatment assignment with the baseline characteristic indicator, 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.

[27] Note that the shaded area around the smoothed conditional mean function in the scatterplots are confidence intervals of the smooth function and do not represent the confidence intervals based on the predicted variance of the GRF algorithm. These are therefore not informative for causal inference, but rather to visualise the estimated out-of-bag CATEs of the GRF algorithm.

Table 8 displays the results of the heterogeneity analysis. We find that treatment effects are significantly higher for children that experienced low levels of parental investment before the start of the program (Column 1). Children in the lowest quartile of the pre-intervention parental investment distribution experienced an increase in skills 0.456 standard deviations larger than children in the top three quartiles of baseline parental investment on average. Similarly, we find that children with low baseline skills benefited significantly more from the program (Column 2). The average treatment effect on infant skill is 0.340 standard deviations higher for children that had infant skills below the median at the start of the intervention compared to those above the median. Lastly, we find no significant differences between children who come from households that are located closer to the Family Planning Offices (column 3). Overall, these results suggest that the parenting intervention was progressive in that it was most effective for children who lagged behind cognitively and came from households where baseline levels of parental investment were initially low[28].

6 Conclusion

This paper reports the results of a randomized trial of a home-based parenting program delivered by cadres employed by China's Family Planning Commission. We find that the program significantly increased infant cognitive skills of children after only six months. There were no significant effects on motor development or social-emotional outcomes. The program also had corresponding positive effects on measures of parental investment and let to a significant increase in parenting skills. Children who lagged behind cognitively and received little parental investment at the onset of the intervention benefited most of the program. These effects occurred despite lackluster compliance with the program which appears to have been driven primarily by a combination of supply-side implementation failures and an unfavorable perception of the FPC by beneficiary households. The program itself, however, had a positive effect on views of the FPC suggesting that public perception may be a less significant obstacle as the program is implemented over time. Efforts to improve supply-side compliance will likely have the greatest impact on improving program effectiveness. These efforts could include measures such as increased monitoring or tying cadre pay to the completion of household visits. Increasing cadre effort on a parenting program may, however, decrease effort on other agency tasks. Efforts to increase supply-side compliance should therefore take this potential cost into account.

Our study faces a number of limitations. First, the study took place in one poor rural area in Northwest

[28] Our main heterogeneity analysis does not examine heterogeneity by trainer characteristics because these are only available for the treatment group. Although we have low power in this limited sample, we present disaggregated treatment effects by trainer characteristics in Appendix Table A5.

China, results may differ in other regions and contexts. While not nationally-representative, the sample chosen for the experiment is reflective of moderately-sized villages in nationally-designated poverty counties that are populated by ethnic Han, places where a program such as this is likely to be targeted in China. Second, children were already over 18 months of age at the start of the trial. It is possible that effects would be larger if children were enrolled at an earlier age and/or the intervention took place over a longer period of time. Finally, we estimate effects only at one point in time at the conclusion of the intervention. Longer-run follow-up of the children in the study will be necessary to determine if the gains we find are lasting or fade out over time. Despite these limitations, our results imply that an ECD program can be effectively delivered through the existing infrastructure of the National Health and Family Planning Commission. Future research should explore alternative interventions to improve ECD outcomes and compare relative cost-effectiveness across alternative delivery models.

References

Almås, I., Attanasio, O., Jalan, J., Oteiza, F., and Vigneri, M. (2018). Using data differently and using different data. *Journal of Development Effectiveness*, 10(4):462–481.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Attanasio, O., Baker-Henningham, H., Bernal, R., Meghir, C., Pineda, D., and Rubio-Codina, M. (2018). Early stimulation: The impacts of a scalable intervention. discussion paper, IFS.

Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C., and Rubio-Codina, M. (2015). Estimating the production function for human capital: Results from a randomized control trial in colombia. 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.

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

Bayley, N. (1969). *Manual for the Bayley scales of infant development*. Psychological Corporation.

Berlinski, S., Schady, N., et al. (2016). *The early years: Child well-being and the role of public policy*. Springer.

Black, M. M. and Dewey, K. G. (2014). Promoting equity through integrated early child development and nutrition interventions. *Annals of the New York Academy of Sciences*, 1308(1):1–10.

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.

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.

Carneiro, P. M. and Heckman, J. J. (2003). Human capital policy.

Carter, M. R., Tjernström, E., and Toledo, P. (2019). Heterogeneous impact dynamics of a rural business development program in nicaragua. *Journal of Development Economics*, 138:77–98.

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

Chan, M. (2013). Linking child survival and child development for health, equity, and sustainable development. *Lancet*, 381(9877):1514–1515.

Chang, S., Zeng, L., Brouwer, I. D., Kok, F. J., and Yan, H. (2013). Effect of iron deficiency anemia in pregnancy on child mental development in rural china. *Pediatrics*, 131(3):e755–e763.

Cirelli, I., Graz, M. B., and Tolsa, J.-F. (2015). Comparison of griffiths-ii and bayley-ii tests for the developmental assessment of high-risk infants. *Infant Behavior and Development*, 41:17–25.

Clarke, D. (2018). Rwolf: Stata module to calculate romano-wolf stepdown p-values for multiple hypothesis testing.

Cunha, F. and Heckman, J. (2007). The technology of skill formation. *American Economic Review*, 97(2):31–47.

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

Davis, J. and Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–50.

Dixit, A. (2002). Incentives and organizations in the public sector: An interpretative review. *Journal of human resources*, pages 696–727.

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.

Gorsuch, R. (1983). Factor analysis. lawrence erlbaum. *Hillsdale, NJ*.

Gorsuch, R. L. (2003). Factor analysis handbook of psychology.

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.

Greenhalgh, S. (1986). Shifts in china's population policy, 1984-86: Views from the central, provincial, and local levels. *Population and Development Review*, pages 491–515.

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., 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.

Hesketh, T., Lu, L., and Xing, Z. W. (2005). The effect of china's one-child family policy after 25 years.

Hesketh, T., Zhou, X., and Wang, Y. (2015). The end of the one-child policy: lasting implications for china. *Jama*, 314(24):2619–2620.

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

Huang, H., Tao, S., Zhang, Y., et al. (1993). Standardization of bayley scales of infant development in shanghai. *Chin J Child Health*, 1(3):158–160.

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.

Li, Q., Yan, H., Zeng, L., Cheng, Y., Liang, W., Dang, S., Wang, Q., and Tsuji, I. (2009). Effects of maternal multimicronutrient supplementation on the mental development of infants in rural western china: follow-up evaluation of a double-blind, randomized, controlled trial. *Pediatrics*, 123(4):e685–e692.

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.

Luiz, D., Barnard, A., Knoesen, N., Kotras, N., Horrocks, S., McAlinden, P., Challis, D., and O'Connell, R. (2006). Griffiths mental development scales: Extended revised. two to eight years. administration manual. *Hogrefe, Oxford, UK*.

Nations, U. (2015). Transforming our world: The 2030 agenda for sustainable development. *New York: United Nations, Department of Economic and Social Affairs*.

Nelson, C. A. and Sheridan, M. A. (2011). Lessons from neuroscience research for understanding causal links between family and neighborhood characteristics and educational outcomes. *Whither opportunity*, pages 27–46.

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.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Romano, J. P. and Wolf, M. (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, 113:38–40.

Squires, J., Bricker, D., and Twombly, E. (2003). The asq: Se user's guide for the ages & stages questionnaires, social-emotional: A parent completed, child-monitoring system for social-emotional behaviors. *Baltimore: Paul H. Brookes Publishing Co*.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

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.

Wilson, J. Q. (1989). *Bureaucracy: What government agencies do and why they do it*. Basic Books.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420–445.

Wu, K. B., Young, M. E., and Cai, J. (2012). *Early child development in China: Breaking the cycle of poverty and improving future competitiveness*. World Bank Publications.

Wu, W., Sheng, D., Shao, J., and Zhao, Z. (2011). Mental and motor development and psychosocial adjustment of chinese children with phenylketonuria. *Journal of paediatrics and child health*, 47(7):441–447.

Yi, S. (1995). Manual of bayley scales of infant development, chinese revision. xiangya school of medicine.

Yi, S., Luo, X., Yang, Z., and Wan, G. (1993). The revising of bayley scales of infant development (bsid) in china. *Chin J Clin Psychol*, 1:71–5.

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.

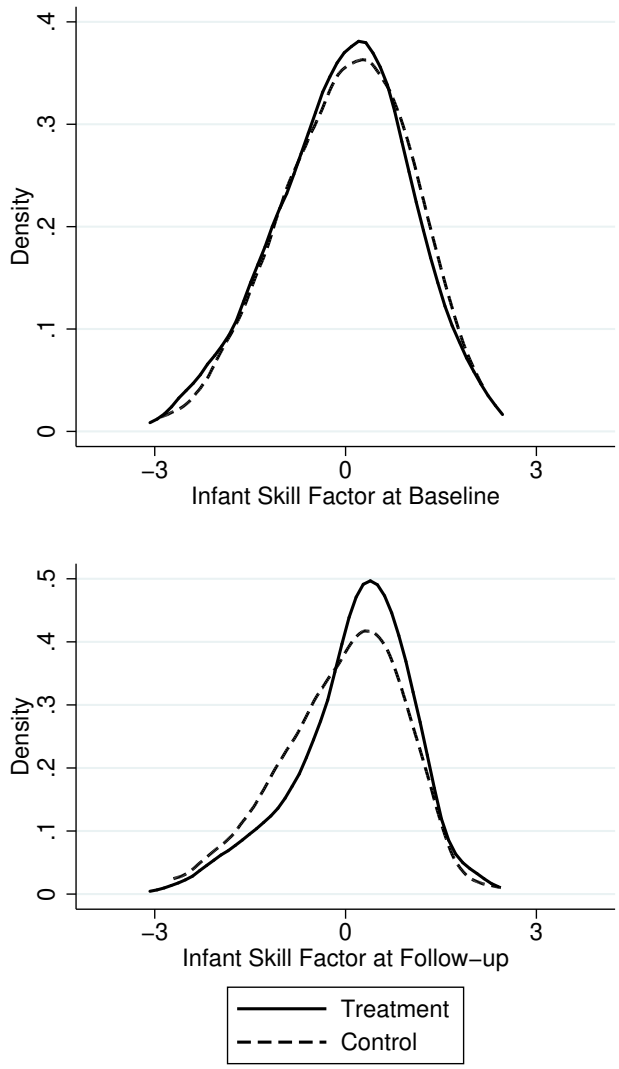
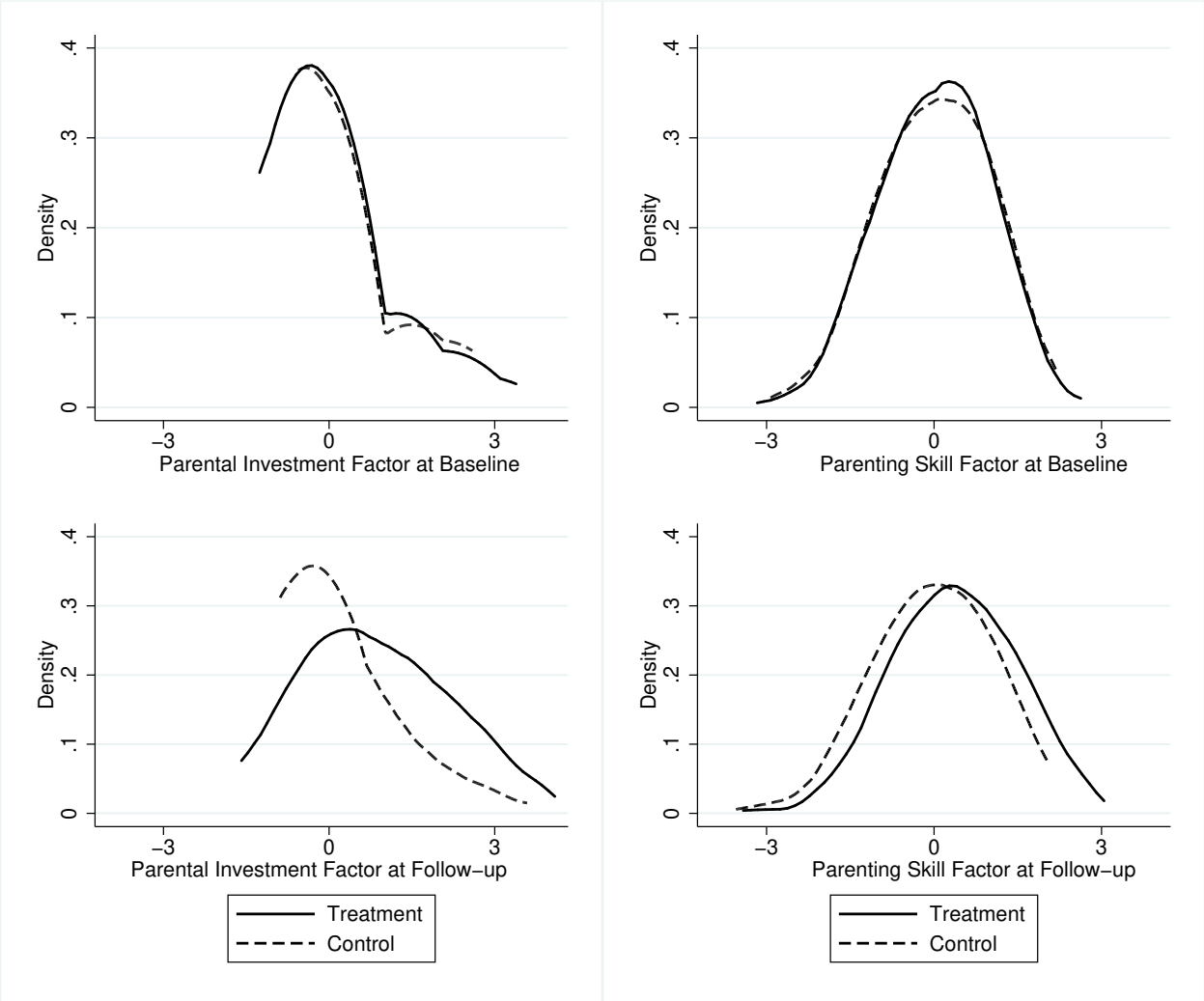


Figure 1: Probability density functions of Bartlett factor scores of Infant Skill are show for baseline and follow-up by treatment assignment. The Kolmogorov-Smirnov (K-S) test of the equality of the infant skill distribution of control and treatment villages cannot be rejected at the 1% significance level (p-value: 0.828) at baseline. At follow-up the K-S test rejects the equality of the two distributions (p-value: 0.029).



(a) Parental Investment Factor

(b) Parenting Skill Factor

Figure 2: Probability density functions of Bartlett factor scores of Parental Investment (a) and Parenting Skill (b) are shown for baseline and follow-up by treatment assignment. The Kolmogorov-Smirnov (K-S) test of the equality of the parental investment and parenting skill distribution of control and treatment villages cannot be rejected at the 1% significance level (p-value:0.973 and 0.889) at baseline. At follow-up the K-S test rejects the equality of the control and treatment distribution for both the parental investment and parenting skill factors (p-value: <0.001 and 0.003).

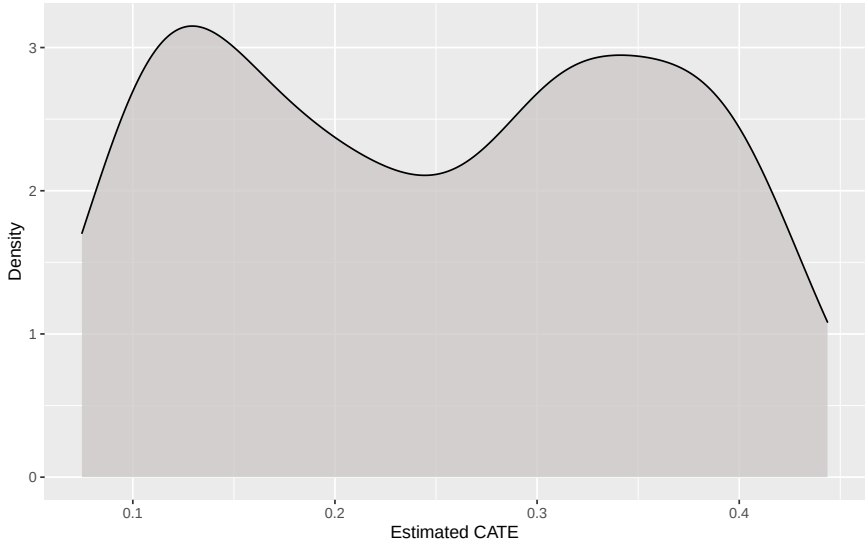


Figure 3: Kernel density function of out-of-bag CATE estimates on Infant Skill from GRF trained algorithm

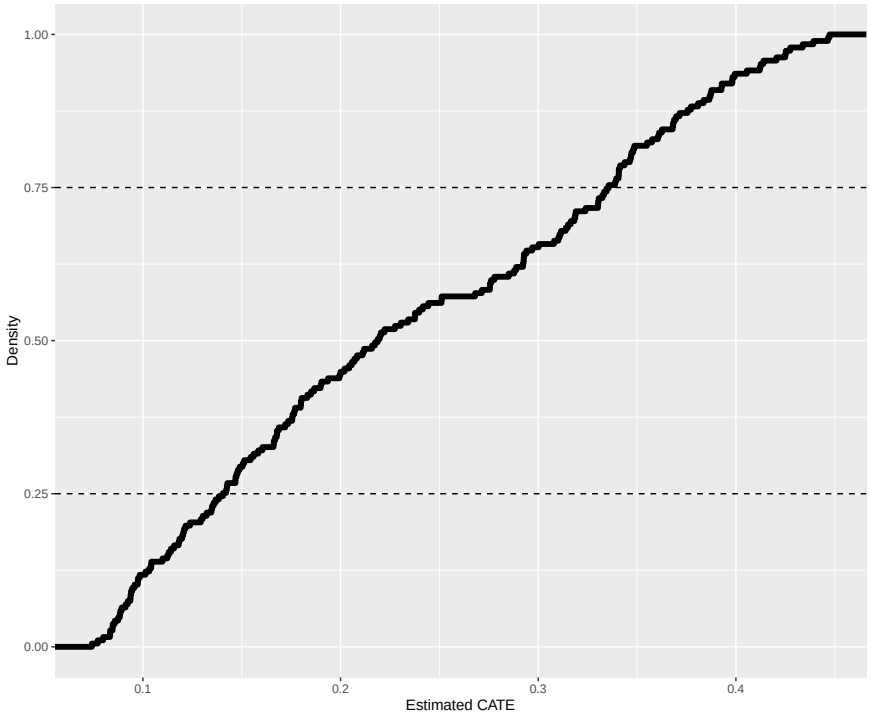


Figure 4: Cumulative distribution function of out-of-bag CATE estimates on Infant Skill from GRF trained algorithm

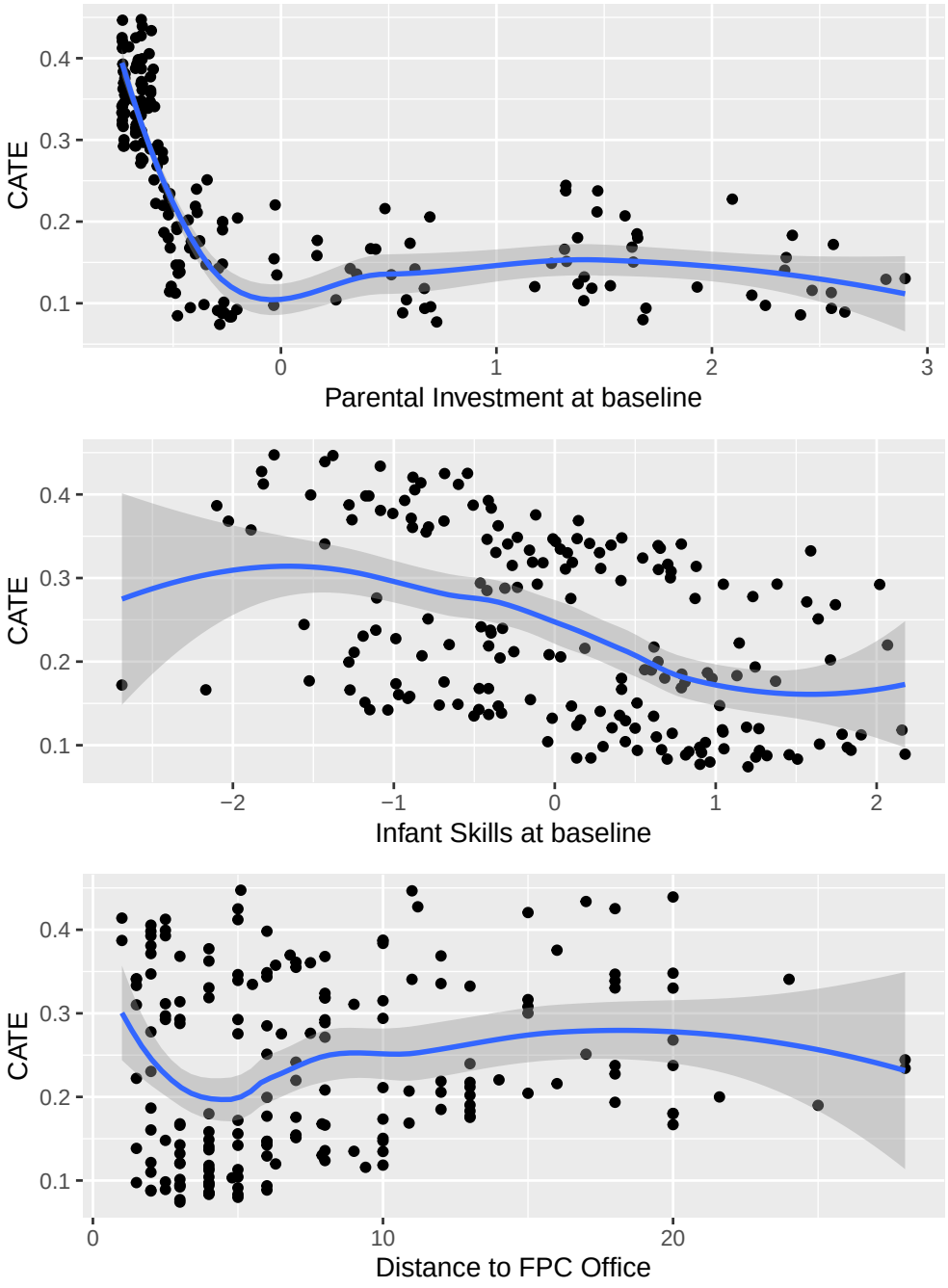


Figure 5: Scatter plots of out-of-bag CATE estimates from GRF trained algorithm along Observable Characteristics

Table 1: Descriptive Statistics and Balance

	(1) Control (N=296)	(2) Treatment (N=212)	(3) p-value
Panel A. Child Characteristics			
(1) Age in months	24.464	24.454	0.975
	(0.198)	(0.220)	
(2) Male	0.449	0.509	0.199
	(0.030)	(0.036)	
(3) Low birth weight	0.041	0.038	0.880
	(0.012)	(0.013)	
(4) First born	0.585	0.612	0.600
	(0.032)	(0.040)	
(5) Ever breastfed	0.847	0.871	0.612
	(0.033)	(0.035)	
(6) Still breastfed ≥ 12 months	0.350	0.387	0.594
	(0.046)	(0.051)	
(7) Anemia (Hb <110 g/L)	0.225	0.272	0.390
	(0.033)	(0.044)	
(8) Days ill past month	4.318	4.548	0.646
	(0.334)	(0.373)	
(9) Cognitive Delay (BSID MDI<80)	0.463	0.389	0.127
	(0.036)	(0.033)	
(10) Motor Delay (BSID PDI<80)	0.123	0.099	0.466
	(0.023)	(0.023)	
(11) Social-Emotional Problems(ASQ:SE>60)	0.250	0.284	0.408
	(0.026)	(0.032)	
Panel B. Household Characteristics			
(1) Social security support recipient	0.279	0.250	0.531
	(0.033)	(0.032)	
(2) Mom at home	0.679	0.621	0.324
	(0.039)	(0.045)	
(3) Caregiver education ≥ 9 years	0.724	0.739	0.732
	(0.026)	(0.035)	
(4) Unfavourable perception of FPC	3.684	3.649	0.784
	(0.091)	(0.091)	
Panel C. Parental Inputs			
(1) Told story to baby yesterday	0.113	0.114	0.986
	(0.020)	(0.024)	
(2) Read book to baby yesterday	0.045	0.043	0.900
	(0.013)	(0.014)	
(3) Sang song to baby yesterday	0.370	0.351	0.695
	(0.030)	(0.038)	
(4) Played with baby yesterday	0.336	0.336	0.988
	(0.028)	(0.033)	
(5) Number of books in household	1.591	1.891	0.422
	(0.236)	(0.290)	

Note: P-values account for clustering at the village level. Unfavourable perception of FPC is measured on a 6-point likert scale.

Table 2: Program Treatment Impact on Infant Skills

	Treatment effect			
	Point estimate	Std. error	P-value	Adjusted P-value
Cohort 1: Below 30 months at follow-up (N=226)				
Bayley: Mental Development Index	0.292**	(0.119)	{0.016}	{0.035}
Bayley: Psychomotor Development Index	-0.024	(0.120)	{0.844}	{0.995}
ASQ: Social-Emotional Problems	-0.010	(0.135)	{0.943}	{0.995}
Cohort 2: Above 30 months at follow-up (N=277)				
Griffith: Performance	0.280**	(0.112)	{0.014}	{0.026}
Griffith: Personal-Social	0.292**	(0.116)	{0.013}	{0.026}
Griffith: Locomotor	-0.018	(0.121)	{0.882}	{0.904}
Griffith: Eye-hand coordination	0.136	(0.126)	{0.281}	{0.465}
ASQ: Social-Emotional Problems	0.118	(0.120)	{0.328}	{0.904}
Infant Skill Factor (N=503)	0.259***	(0.081)	{0.002}	

Note: In all regressions we control for strata (county) fixed effects, previous nutrition assignment status and baseline developmental outcomes. In the pooled factor regression we additionally control for cohort fixed effects. All development outcomes are non-parametrically standardized for each age-month group. The Griffith language subscale is omitted in the analysis for the older cohort as receptive and expressive language skills are not explicitly tested by the BSID I and we want comparable measures of infant skills across both age groups. We find a positive but insignificant treatment effect on the Griffith language subscale (point estimate: 0.023 and std. error: 0.107). All standard errors are clustered at the village level. Adjusted P-values are calculated using the Romano Wolf (2005) stepdown-procedure to control for the family-wise error rate (FWER). Significance levels based on adjusted P-values are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Program Treatment Impacts on Parenting Skills and Parental Investment

	Treatment effect			
	Point estimate	Std. error	P-value	Adjusted P-value
Panel A. Parenting Skills at follow-up (N=475)				
Parent feels duty to help baby understand the world	0.074	(0.079)	{0.348}	{0.751}
Parent knows how to play with baby	0.062	(0.089)	{0.478}	{0.703}
Parent knows how to read stories to baby	0.304***	(0.087)	{0.001}	{0.002}
Parent finds it important to play with baby	0.058	(0.092)	{0.528}	{0.703}
Parent finds it important to read stories to baby	0.304***	(0.088)	{0.001}	{0.002}
Parent finds it difficult to communicate with baby	0.053	(0.099)	{0.592}	{0.751}
Parent feels nervous when caring for baby	-0.144	(0.091)	{0.117}	{0.389}
Parenting Skill Factor	0.323***	(0.091)	{0.001}	
Panel B. Parental Investment at follow-up (N=475)				
Number of books in hh for reading to baby	0.291***	(0.091)	{0.002}	{0.001}
Number of times per week family reads to baby	0.897***	(0.116)	{<0.001}	{0.001}
Number of times per week family sings to baby	0.362***	(0.085)	{<0.001}	{0.001}
Number of times per week family goes out with baby	-0.042	(0.094)	{0.658}	{0.951}
Number of hours per day baby spends watching tv	0.048	(0.244)	{0.844}	{0.991}
Number of hours per day baby plays by itself	0.125	(0.108)	{0.249}	{0.848}
Parental Investment Factor	0.825***	(0.107)	{<0.001}	

Note: In all regressions we control for strata (county) fixed effects, previous nutrition assignment status and baseline parental skills or investment measures. In the pooled factor regressions we additionally control for cohort fixed effects. All outcomes are standardized by the distribution of the control group. Parenting skill outcomes are measured on a 7-point likert scale. Number of times per week family reads, sings or goes out with baby are measured on a 4-point likert scale. All standard errors are clustered at the village level. All standard errors are clustered at the village level. Adjusted P-values are calculated using the Romano Wolf (2005) stepdown-procedure to control for the familywise error rate (FWER). Significance levels based on adjusted P-values are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4a: Determinants of Compliance

	(1) HH Visits	(2) HH Visits	(3) HH Visits	(4) HH Visits	(5) HH Visits
Male	1.599*	1.965**	1.935**	1.849**	1.398*
	(0.823)	(0.849)	(0.841)	(0.853)	(0.831)
Age in months	-0.083	-0.040	-0.038	0.005	-0.038
	(0.118)	(0.115)	(0.116)	(0.123)	(0.100)
Cognitive Delay (BSID: MDI<80)	-1.541*	-1.691**	-1.526*	-1.548*	-1.181
	(0.851)	(0.840)	(0.834)	(0.827)	(0.746)
Motor Delay (BSID: PDI<80)	-1.130	-1.573	-1.897*	-1.714	-0.556
	(1.201)	(1.089)	(1.072)	(1.113)	(1.026)
Social-Emotional Problems (ASQ: SE>60)	0.110	0.663	0.930	0.662	0.946
	(0.972)	(0.837)	(0.842)	(0.853)	(0.844)
Number of days ill	0.085	0.037	0.045	0.030	-0.045
	(0.132)	(0.131)	(0.130)	(0.129)	(0.126)
Mom home > 2 years		0.652	0.596	0.911	0.741
		(1.067)	(1.021)	(0.984)	(0.865)
Maternal education > 9 year		1.136	0.973	1.048	0.534
		(0.961)	(0.926)	(0.886)	(0.974)
Social security support recipient		-1.582	-1.916*	-1.821*	-1.412
		(0.999)	(0.985)	(1.036)	(1.069)
Distance to FPC office		-0.326***	-0.331***	-0.339***	-0.334***
		(0.116)	(0.115)	(0.118)	(0.115)
Unfavourable perception of FPC			-1.467***	-1.562***	-1.839***
			(0.518)	(0.528)	(0.506)
Trainer is male				-1.214	-1.296
				(1.400)	(1.374)
Trainer work experience FPC				0.144	0.146
				(0.110)	(0.113)
Trainer has bachelor degree				0.045	-0.490
				(1.417)	(1.107)
County FE	No	No	No	No	Yes
Observations	211	211	211	211	211
R^2	0.04	0.13	0.16	0.18	0.26

Note: Unfavorable perception of Family Planning Commission (FPC) is measured on a 5-point likert scale. Trainer work experience is measured by the number of years worked as a cadre for the FPC. All Standard errors are clustered at the village level. Significance levels are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4b: Determinants of Compliance

	(1) Distance to FPC	(2) Distance to FPC	(3) Unfavourable perception of FPC	(4) Unfavourable perception of FPC
Male	0.784	0.757	-0.024	-0.043
	(0.583)	(0.546)	(0.109)	(0.107)
Age in months	0.171*	0.149	0.004	0.003
	(0.100)	(0.103)	(0.017)	(0.017)
Cognitive Delay (BSID: MDI<80)	-0.346	-0.289	0.112	0.128
	(0.728)	(0.631)	(0.112)	(0.118)
Motor Delay (BSID: PDI<80)	-1.201	-1.295	-0.210	-0.146
	(1.020)	(1.042)	(0.140)	(0.136)
Social-Emotional Problems (ASQ: SE>60)	0.781	1.069	0.158	0.136
	(0.822)	(0.859)	(0.134)	(0.127)
Number of days ill	-0.121	-0.078	0.004	-0.004
	(0.087)	(0.090)	(0.013)	(0.015)
Mom home > 2 years	0.956	0.411	-0.019	0.023
	(0.879)	(0.823)	(0.107)	(0.097)
Maternal education > 9 year	0.630	1.080	-0.113	-0.167
	(0.814)	(0.814)	(0.123)	(0.120)
Social security support recipient	0.000	0.232	-0.216**	-0.211**
	(0.779)	(0.754)	(0.095)	(0.102)
Trainer is male	-2.046	-1.997	-0.075	-0.088
	(1.316)	(1.298)	(0.131)	(0.126)
Trainer work experience FPC	-0.084	-0.124	0.009	0.014
	(0.083)	(0.087)	(0.007)	(0.008)
Trainer has bachelor degree	-2.480**	-3.074***	0.104	0.133
	(1.028)	(1.037)	(0.131)	(0.129)
County FE	No	Yes	No	Yes
Observations	211	211	211	211
R^2	0.13	0.17	0.06	0.09

Note: Standard errors are clustered at the village level. Significance levels are as follows: * $p < 0.1$, ** $p < 0.05$

Table 5: Average Treatment Effect on Perception of Family Planning Commission

	(1) Unfavorable Perception FPC
Treatment	-0.332**
	(0.134)
Observations	512
R^2	0.06
Control mean	3.80

Note: We control for strata (county) fixed effects, cohort fixed effects, and previous nutrition assignment status. Perception of FPC is measured on a 6-point likert scale. Standard errors are clustered at the village level and reported in parentheses. Significance levels are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Dose-Response Relationships

	(1) Infant Skill	(2) Infant Skill	(3) Parenting Skill	(4) Parenting Skill	(5) Parental Inv.	(6) Parental Inv.
Number of HH Visits	0.014***	0.052	0.019***	0.056	0.049***	0.073
	(0.005)	(0.037)	(0.005)	(0.047)	(0.006)	(0.055)
Number of HH Visits^2		-0.002		-0.002		-0.002
		(0.002)		(0.003)		(0.003)
Observations	503	503	475	475	475	475
R^2	0.22	0.22	0.08	0.09	0.25	0.25

Note: Column (1), (3) and (5) give control function estimates of the treatment effect of one household visit on the factor outcomes of interest, assuming a linear relationship between the number of household visits and the factor outcomes up to 24 household visits. Column (2), (4) and (6) give control function estimates of the treatment effect of one household visit, assuming a concave relationship. Residuals used in the control function estimation are derived from regressing the number of household visits on treatment status, distance to the FPC office and the interaction of the distance measure with treatment assignment. Estimates of the fist stage regression can be found in Appendix Table A4. F-test of joint significance of the excluded instruments gives a p-value <0.001. In all regressions we control for baseline latent factors, strata(county) fixed effects, cohort fixed effects and previous nutrition assignment status. All standard errors are clustered at the village level. Significance levels are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Baseline Characteristics used in GRF Analysis Ranked by Variable Importance

Baseline Characteristics	Variable Importance
Parental investment	27.16%
Infant skills	16.73%
Distance to FPC Office	12.51%
Number of days ill	11.27%
Parenting skills	9.65%
Household Assets	7.75%
Mother at home	7.31%
Caregiver education $\geq$ 9 years	2.43%
Male	1.78%
Unfavourable perception of FPC at county level	1.33%
Social security support recipient	1.07%
Unfavourable perception of FPC at village level	1.02%

Note: Variable importance is the frequency with which each observable baseline characteristic is used as a splitting variable in the Generalized Random Forest (GRF) algorithm.

Table 8: Heterogeneous Treatment Effects on Cognitive Development

	(1) Infant Skill	(2) Infant Skill	(3) Infant Skill
Treatment	0.072	0.065	0.259***
	(0.104)	(0.096)	(0.096)
First quartile of parental investment * treatment	0.456*		
	(0.238)		
First quartile of parental investment	-0.398*		
	(0.206)		
Below median infant skill * treatment		0.340**	
		(0.153)	
Below median infant skill		-0.725***	
		(0.108)	
First quartile of distance to FPC * treatment			-0.157
			(0.196)
First quartile of distance to FPC			-0.011
			(0.144)
Observations	473	508	508
R^2	0.07	0.13	0.05

Note: In all regressions we control for strata (county) fixed effects and cohort fixed effects. Infant skill outcomes are non-parametrically standardized for each age-month group. All standard errors are clustered at the village level. Significance levels based on P-values are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Appendix A

Table A1: Trainer Summary Statistics (N=69)

Variable	Mean	Std. Dev.
Male	0.623	0.488
Age	34.246	5.984
Married	0.899	0.304
Has child	0.855	0.355
Age of youngest child	7.134	6.286
Has bachelor degree	0.290	0.457
Monthly Salary (RMB)	3238.159	496.749
Work experience FPC (years)	12.116	7.118

Table A2: Descriptive Statistics and Balance

	(1) Control (N=296)	(2) Treatment (N=212)	(3) Spillover (N=79)	(4) P-value Control vs. Treatment	(5) P-value Control vs. Spillover	(6) P-value Treatment vs. Spillover
Panel A. Child Characteristics						
(1) Age in months	24.468	24.454	24.379	0.962	0.814	0.842
	(0.199)	(0.220)	(0.328)			
(2) Male	0.450	0.509	0.582	0.211	0.020	0.152
	(0.030)	(0.036)	(0.047)			
(3) Low birth weight	0.041	0.038	0.051	0.874	0.749	0.697
	(0.012)	(0.013)	(0.029)			
(4) First born	0.583	0.612	0.658	0.581	0.246	0.524
	(0.032)	(0.040)	(0.056)			
(5) Ever breastfed	0.846	0.871	0.872	0.597	0.690	0.989
	(0.033)	(0.035)	(0.057)			
(6) Still breastfed ≥ 12 months	0.346	0.387	0.333	0.545	0.891	0.557
	(0.046)	(0.051)	(0.077)			
(7) Anemia (Hb <110 g/L)	0.226	0.272	0.164	0.399	0.283	0.102
	(0.033)	(0.044)	(0.048)			
(8) Days ill past month	4.323	4.548	4.768	0.653	0.618	0.813
	(0.335)	(0.373)	(0.835)			
(9) Cognitive Delay (BSID MDI<80)	0.464	0.389	0.364	0.118	0.236	0.760
	(0.036)	(0.033)	(0.078)			
(10) Motor Delay (BSID PDI<80)	0.124	0.099	0.127	0.459	0.950	0.642
	(0.023)	(0.023)	(0.055)			
(11) Social-Emotional Problems(ASQ:SE>60)	0.251	0.284	0.321	0.421	0.238	0.580
	(0.026)	(0.032)	(0.054)			
Panel B. Household Characteristics						
(1) Social security support recipient	0.280	0.250	0.291	0.519	0.865	0.504
	(0.033)	(0.032)	(0.057)			
(2) Mom at home	0.682	0.621	0.661	0.305	0.771	0.589
	(0.039)	(0.045)	(0.061)			
(3) Caregiver education ≥ 9 years	0.724	0.739	0.782	0.716	0.239	0.339
	(0.026)	(0.035)	(0.042)			
(4) Unfavourable perception of FPC	3.676	3.649	3.745	0.838	0.701	0.596
	(0.091)	(0.091)	(0.159)			
Panel C. Parental Inputs						
(1) Told story to baby yesterday	0.114	0.114	0.089	0.997	0.567	0.593
	(0.020)	(0.024)	(0.038)			
(2) Read book to baby yesterday	0.046	0.043	0.018	0.893	0.214	0.288
	(0.013)	(0.014)	(0.018)			
(3) Sang song to baby yesterday	0.367	0.351	0.464	0.731	0.273	0.182
	(0.030)	(0.038)	(0.084)			
(4) Played with baby yesterday	0.333	0.336	0.375	0.942	0.537	0.583
	(0.028)	(0.033)	(0.062)			
(5) Number of books in household	1.597	1.891	2.304	0.432	0.300	0.548
	(0.236)	(0.290)	(0.644)			

Note: P-values account for clustering at the village level.

Table A3: Average Treatment Effects on Infant Skills, Parenting Skills and Parental Investment of Non-treated Children in Treatment Villages (N=79)

	Treatment effect	
	Point estimate	Std. error
Infant Skill Factor (N=369)	0.119	(0.107)
Parenting Skill Factor (N=319)	-0.055	(0.150)
Parental Investment Factor (N=319)	-0.045	(0.154)

Note: In all regressions we control for strata (county) fixed effects, cohort fixed effects, previous nutrition assignment status and baseline latent factors. All standard errors are clustered at the village level. Significance levels are as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: First Stage of Dose-Response Relationship

	(1)	(2)	(3)
Excluded Instruments			
Treatment	18.774***	18.756***	18.782***
	(1.101)	(1.103)	(1.092)
Distance to FPC office	-0.002	-0.005	-0.002
	(0.019)	(0.021)	(0.021)
Distance to FPC office * Treatment	-0.294**	-0.286**	-0.292**
	(0.115)	(0.117)	(0.116)
Lagged Outcome Variables			
Bayley: Mental Development Index	-0.219		
	(0.226)		
Bayley: Psychomotor Develoment Index	0.428**		
	(0.214)		
ASQ: Social-Emotional Problems	0.497**		
	(0.236)		
Parenting skill		0.001	
		(0.226)	
Parental investment			-0.290
			(0.177)
Observations	507	475	475
R^2	0.84	0.83	0.83
F-stat excluded instruments	210.50	209.98	212.87

Note: In all regressions we control for strata (county) fixed effects, cohort fixed effects and previous nutrition assignment status. All standard errors are clustered at the village level. $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

Table A5: Heterogeneity of Program Impact by Trainer Characteristics

		Infant Skill (N=503)
Trainer Gender		
	Male	0.204**
		(0.101)
	Female	0.322*
		(0.092)
	P-value test equality	0.289
Trainer Age		
	Below 33 years	0.299***
		(0.099)
	33 years and above	0.206**
		(0.100)
	P-value test equality	0.420
Trainer Experience		
	Below 12 years	0.297***
		(0.101)
	12 years and above	0.205**
		(0.097)
	P-value test equality	0.424
Trainer Education		
	Below Bachelor degree	0.161*
		0.096
	Bachelor degree	0.289***
		0.094
	P-value test equality	0.237

Note: In all regressions we control for strata (county) fixed effects, cohort fixed effects, previous nutrition assignment status and baseline latent factors. All standard errors are clustered at the village level. $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

Appendix B: Measurement System

In this appendix we provide further detail about the measurement system relating observed measures to the latent factors of infant skill, parenting skill and parental investment used in the analysis. We follow the psychometric literature (Gorsuch, 1983, 2003) and recent economic research in early childhood development (Heckman et al., 2013; Attanasio et al., 2015) and aim to develop a measurment system with dedicated measures which only proxy one latent factor. First, we provide results of the exploratory factor analysis (EFA) which informed the specification of our dedicated measurement system. Next, we present estimates of the dedicated measurement system.

B.1 Exploratory Factor Analysis

Exploratory factor analysis is used to select the number of latent factors that need to be extracted from all the measures we have on infant skill, parenting skill and parental investment. Once the number of latent factors is determined for each of these three dimensions we estimate factor loadings and allocate measures to factors. Measurments that have weak loadings or cross-load on multiple factors are discarded in order to achieve a dedicated measurement system that makes the interpretation of the latent factors transparent. We base the EFA on baseline measures collected before the parenting intervention started.

Many methods are developed in the literature to select the number of factors and we use two of the most widely used methods to guides the factor selection process: Horns's parallel analysis (Horn, 1965) and Cattell's scree plot (Cattell, 1966). Table B1 shows the number of factors both methods suggest that should be extracted from the measures we have on infant skills, parenting skills and parental investment at baseline.

Table B1: Exploratory factor analysis to determine the number of latent factors

	Cattell's scree plot	Horn's parallel analysis
Measured dimensions		
Infant skill at baseline	1	1
Parenting skill at baseline	1	2
Parental investment at baseline	1	2

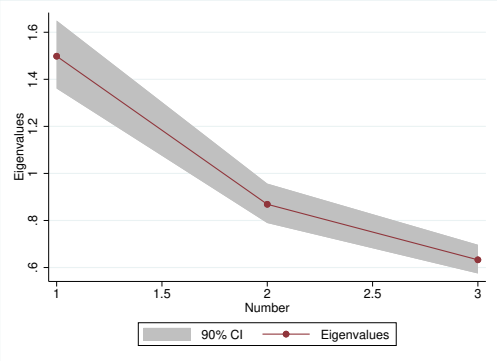


Figure B1: Scree Plot of Eigenvalues of PCA for Infant Skills

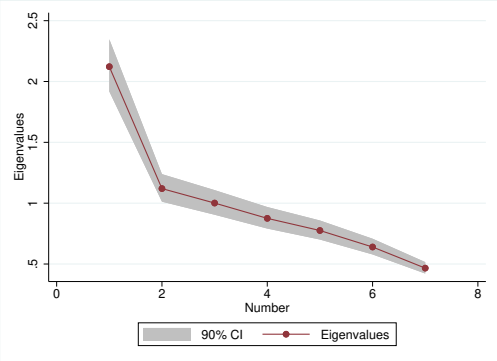


Figure B2: Scree Plot of Eigenvalues of PCA for Parenting Skills

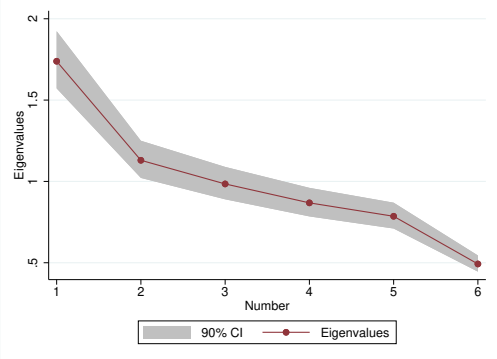


Figure B3: Scree Plot of Eigenvalues of PCA for Parental Investment

For our measures on infant skill both methods indicate that we extract one factor. For parenting skill and parental investment the analysis suggest we should extract 1 or 2 factors. We next proceed with estimating factor loadings to allocate measures to factors and discard measures that proxy the latent factor only weakly or cross-load on factors. for the two-factor models we use the quartimin rotation method in this second step of the EFA which rotates estimated factor loadings in order to identify measures that strongly load on one factor. This allows us to choose the best measures for the dedicated measurement system. Table B2 reports estimated factor loadings for each of the infant skill measures at baseline.

Table B2: Estimated factor loadings on infant skills at baseline

	First Factor
One-Factor Model	
Bayley: Mental Development Index	0.530
Bayley: Psychomotor Development Index	0.478
ASQ: Social-Emotional Problems	-0.340

Both the Bayley Mental Development and Psychomotor Development index load positively and strongly on the latent factor. The social-emotional problem index from the Ages and Stages Questionnaire (ASQ) loads negatively on the latent factor, which gives us confidence we are indeed measuring infant skills as higher values of the ASQ indicate developmental problems. Given that the ASQ is a carer-reported instrument to measure child social and emotional development it suffers more from measurement error than the Bayley indexes which are assessed by trained personnel. For our baseline ASQ measure we have therefore taken the average ASQ score of 3 assessment periods prior to the intervention in an attempt to mitigate the measurement error problem.[29]

Table B3 reports the estimated factor loadings for the measures of parenting skills that were collected at baseline. We present both results for a one-factor and two-factor model given that the Horn's parallel analysis (Horn, 1965) suggested a second factor could be extracted from the measures. The pattern of factor loadings in both the one- and two factor model clearly support one grouping of measures. The first five measures in Table 3 load strongly on the first factor and proxy for parenting skills. On the other hand, the factor loadings on the level of difficulty in communication care-givers experience towards their offspring and their feelings of nervousness about child-rearing do not load clearly on either factor. We therefore exclude these two measures as they are not good proxy measures for our dedicated measurement system. In

[29] Given that the treatment assignment for the parenting intervention evaluated in this study was stratified on the arms of an earlier micro-nutrient trial we have multiple carer-reported ASQ measures.

the final measurement system we hence retain the first five measures (highlighted in grey in Table 3) both at baseline and follow-up to proxy for the factor we interpret as parenting skill.

Table B3: Estimated factor loadings on parenting skills at baseline

	First Factor	Second Factor
One-Factor Model		
Parent feels duty to help baby understand the	0.414	
Parent knows how to play with baby	0.511	
Parent knows how to read stories to baby	0.499	
Parent finds it important to play with baby	0.527	
Parent finds it important to read stories to baby	0.563	
Parent finds it difficult to communicate with baby	-0.129	
Parent feels nervous when caring for baby	-0.210	
Two-Factor Model		
Parent feels duty to help baby understand the	0.397	0.287
Parent knows how to play with baby	0.504	0.139
Parent knows how to read stories to baby	0.513	-0.219
Parent finds it important to play with baby	0.513	0.230
Parent finds it important to read stories to baby	0.573	-0.146
Parent finds it difficult to communicate with baby	-0.141	0.200
Parent feels nervous when caring for baby	-0.214	0.053

Estimated factor loadings on measures of parental investment at baseline are reported in Table B4. We find that the number of children's books in the household and the time spend reading and singing with the child at baseline load strongly on the fist factor. The measures capturing the time the child spends playing alone or watching tv and the time the child spends in outdoor activities with the caregiver do not load clearly on any of the two factors and are therefore discarded from the dedicated measurement system. For both the baseline and follow-up factor proxying parental investment we hence retain the three first measures (as highlighted in grey) for the dedicated measurement system.

B.2 Estimates of the Dedicated Measurement System

Table B5 reports the estimates of the dedicated measurement system at baseline and follow-up. The first column reports the factor loadings for each of the dedicated measures. We normalized the factor loadings

Table B4: Estimated factor loadings on parental investment at baseline

	First Factor	Second Factor
One-Factor Model		
Number of books in hh for reading to baby	0.453	
Number of times per week family reads to baby	0.648	
Number of times per week family sings to baby	0.526	
Number of times per week family goes out with baby	0.220	
Number of hours per day baby spends watching tv	0.067	
Number of hours per day baby plays by itself	0.030	
Two-Factor Model		
Number of books in hh for reading to baby	0.453	0.043
Number of times per week family reads to baby	0.648	-0.015
Number of times per week family sings to baby	0.526	0.011
Number of times per week family goes out with baby	0.218	-0.202
Number of hours per day baby spends watching tv	0.068	0.175
Number of hours per day baby plays by itself	0.032	0.291

of the first measure at baseline and follow-up to one. Hence, at baseline the scale of the latent infant skill factor is determined by the Bayley Mental Development Index. At follow-up, the scale of the latent infant factor is determined by the Bayley Mental Development Index for the younger cohort, and by the Griffith Performance Index for the older age cohort. Similarly, the scale of both the parenting skill factor and the parental investment factor at baseline and follow-up are determined by the first measure. The second column of Table B5 shows estimates for how much of the variance is driven by signal relative to noise. The signal-to-noise ratios for the m^{th} measure of child development is calculated as:

$$S_m^{\theta} = \frac{\lambda_m^2 Var(\theta)}{\lambda_m^2 Var(\theta) + Var(\delta_m)}$$

As shown in Table B5, most measures are far away from having 100 % of their variance accounted for by signal which highlights the usefulness of the latent factor approach when modelling parental investment and early skill formation. The survey measurement error typically present in these variables would risk to lead to severely attenuated coefficients in the absence of a dedicated measurement approach. We find that this is specifically the case for the *ASQ: Social-Emotional Problems* index which has a relatively low signal-to-noise ratio compared to the Bayley and Griffith indexes of child development. Given that the

ASQ is a caregiver-reported instrument to measure child social and emotional development it suffers more from measurement error than the Bayley and Griffith indexes which are assessed by trained personnel (Johnston et al., 2014). For our baseline ASQ measure we have therefore taken the average ASQ score of 3 assessment periods prior to the intervention in an attempt to mitigate the measurement error problem and as can be seen in Table B5 the signal-to-noise ratio for the ASQ measure is indeed better at baseline than at follow up. As Cunha et al. (2010) show, the distribution of measurement error and the latent factor distribution are non-parametrically identified as long as we have at least 3 measures with nonzero factor loading corresponding to each latent factor. Hence, we keep the ASQ measure in the dedicated measurement system for infant skills despite the relatively low singal-to-noise ratio at follow-up.

Table B5: Dedicated Measurement System

Latent Factor	Measurement	factor loading	% Signal
Infant Skill Factor at baseline			
	Bayley: Mental Development Index	1	0.560
	Bayley: Psychomotor Development Index	0.613	0.222
	ASQ: Social-Emotional Problems	-0.455	0.100
Infant Skill Factor at follow-up			
Age-Cohort 1	Bayley: Mental Development Index	1	0.435
	Bayley: Psychomotor Development Index	0.749	0.249
	ASQ: Social-Emotional Problems	-0.287	0.039
Age-Cohort 2	Griffith: Performance	1	0.347
	Griffith: Personal-Social	1.142	0.419
	Griffith: Locomotor	1.162	0.467
	Griffith: Eye-hand coordination	1.022	0.338
	ASQ: Social-Emotional Problems	-0.320	0.034
Parenting Skill Factor at baseline			
	Parent feels duty to help baby understand the world	1	0.171
	Parent knows how to play with baby	1.595	0.251
	Parent knows how to read stories to baby	1.798	0.239
	Parent finds it important to play with baby	1.193	0.323
	Parent finds it important to read stories to baby	1.579	0.347
Parenting Skill Factor at follow-up			
	Parent feels duty to help baby understand the world	1	0.072
	Parent knows how to play with baby	2.803	0.214
	Parent knows how to read stories to baby	4.337	0.388
	Parent finds it important to play with baby	1.598	0.168
	Parent finds it important to read stories to baby	2.915	0.350
Parental Investment Factor at baseline			
	Number of books in hh for reading to baby	1	0.154
	Number of times per week family reads to baby	0.583	0.971
	Number of times per week family sings to baby	0.328	0.190
Parental Investment Factor at follow-up			
	Number of books in hh for reading to baby	1	0.104
	Number of times per week family reads to baby	0.494	0.622
	Number of times per week family sings to baby	0.418	0.290

Note: Table shows dedicated measurement system. For each measure factor loadings are shown as well as the fraction of the variance in each measure that is explained by the variance in signal.