Dynamic complementarity in early capability formation: Evidence from a cluster-randomized parenting experiment in rural China

By DORIEN EMMERS, RENFU LUO, SCOTT ROZELLE AND SEAN SYLVIA*

We use a cluster-randomized controlled trial to investigate how a two-year, home-based parenting training program drives the capability formation of 1-year-olds (N = 449) in rural China. We find that biweekly parenting training in child psychosocial stimulation and health promotion delivered by community health workers raises child skill development, on average, by 0.18 SD and 0.33 SD after one and two years, respectively. Analysis of impact heterogeneity between subgroups of children with high versus low early capabilities suggests (1) dynamic self-complementarity between first-year skills and two-year productivity of skill investment and (2) dynamic cross-complementarity between first-year health and two-year productivity of skill investment.

* Emmers: Stanford University and Catholic University of Leuven (email: dorien.emmers@kuleuven.be); Luo: Peking University (email: luorf.ccap@pku.edu.cn); Rozelle: Stanford University (email: rozelle@stanford.edu); and Sylvia: University of North Carolina at Chapel Hill (email: sean_sylvia@unc.edu). Approval for the study was obtained from the provincial offices of the National Health Commission in Yunnan and Hebei in 2017 and from the ethics committee of Stanford University (USA, No. 25734) on October 26, 2012, and renewed annually. The trial is registered with the American Economic Association's registry for randomized controlled trials (AEA RCT Registry), No. AEARCTR-0009789 and the International Standard Randomized Controlled Trial Number (ISRCTN) registry, number ISRCTN72589193. The authors gratefully acknowledge funding from the National Science Foundation of China (Grant No. 71873008), the Data Center of Management Science, National Natural Science Foundation, Peking University (Grant No. 2017KEY04), and Save the Children Hong Kong. Emmers acknowledges support from KU Leuven, long-term structural Methusalem funding from the Flemish Government (FWO) under Excellence of Science (EOS) Project No. G0G4318N (EOS ID 30784531). The authors have nothing to disclose.

The first several years of life are a sensitive and critical period for human capital formation because neural and behavioral plasticity is highest during this period (Almond and Currie 2011; Cunha and Heckman 2007; Knudsen et al. 2006). Skill and health formation during early childhood have been linked not only to cognitive, social-emotional functioning, and health in adulthood but also to more distal outcomes, including wages and participation in crime (Gertler et al. 2014; Heckman 2012; Knudsen et al. 2006; Walker et al. 2021). These findings have salient implications for low- and middle-income countries (LMICs), where, according to estimates based on proxy measures of risk factors for child development, such as poor home environments and nutrition, approximately 250 million children (or 43%) below the age of five are at risk of early skill and health deficits (Black et al. 2017). In addition to individual benefits, the potential consequences and scale of these capability deficits imply that effective policies and interventions to promote early childhood development (ECD) in LMICs have large social benefits through increased accumulation of human capital, a key driver of economic growth (Solow 1956); reduced intergenerational transmission of poverty and increased social mobility (Heckman and Mosso 2014); reduced social inequality (Heckman 2013); and other society-level effects.

Parenting training programs that promote child psychosocial stimulation consistently result in positive impacts on caregivers' engagement in cognitively stimulating parenting practices and early cognitive capabilities of young children in LMICs (Rao et al. 2017). Nevertheless, many questions related to the optimal design of parenting programs remain to be investigated. A number of prominent studies put integration of multiple program components that target diverse child capabilities (e.g., health, cognitive skills) forward as a promising means to improve the cost effectiveness of programs (see, e.g., Alderman et al. 2014; Yousafzai and Aboud 2014). The seminal theoretical work of Cunha and Heckman (2007) on the process of capability formation provides support for this argument. According to

Cunha and Heckman, integration of program components can stimulate dynamic complementarities in the process of capability formation. Dynamic complementarity implies that capabilities produced in one period can lead to higher productivity of investments in subsequent periods. Dynamic complementarities, together with dynamic productivity effects, can produce multiplier effects through which capabilities beget capabilities, which has important implications for the optimal timing of intervention programs due to the age-dependent economic returns on investments (Alderman et al. 2014). To date, however, there is limited empirical evidence on how integration of program components can lead to synergistic dynamic effects in capability formation (Hurley et al 2016).

In this paper, we evaluate the impacts of a two-year home visitation program that targets skill development and health. A cluster-randomized controlled trial was carried out between October 2015 and October 2017 in rural areas of two provinces in China. We administered the cognition, language, motor, and social-emotional subscales of the third edition of the Bayley Scales of Infant and Toddler Development (Bayley-III) to assess child skill development. We measured hemoglobin values and disease incidence to evaluate impacts on child health. We find that biweekly parenting training on child psychosocial stimulation and health promotion delivered by community health workers improves aggregated Bayley-III skill development scores after one year and two years by 0.18 SD and 0.33 SD, respectively. Aggregated child health scores improve, on average, by 0.25 SD after one year, but this impact fades out during the second year, when the control group health catches up naturally with that of the treatment group. In addition, we observe that these improvements in skill and health capabilities concur with increases in mediating parental inputs, such as parental investment (i.e., in child psychosocial stimulation and health) and parenting-related knowledge.

We then use the experiment to explore dynamic complementarities in the formation of early capabilities and determine the implications for optimal program

design. We investigate heterogeneity in treatment effects between subgroups with high or low initial capabilities (i.e., skills or health) to explore dynamic complementarities in the skill-building process. In line with the work of Aizer and Cunha (2012), we assume that the interaction effect between the initial stock of capabilities and randomized treatment assignment (i.e., a source of exogenous variation in parental investment) captures the presence of dynamic complementarity between early capabilities and investments in the production of later capabilities. To explore dynamic complementarities at age 1 and age 2, we define subgroups based on pretreatment assessments of capabilities (at age 1) and predictions of the attained capabilities at the end of the first intervention year (at age 2). As suggested by Abadie et al. (2018), we predict skills and health at the end of the first year based on pretreatment covariates, whereby we use a random forest machine learning algorithm—the prediction algorithm with the lowest *K*-fold cross-validation mean squared prediction error (MSPE; Breiman 2001). The subgroup heterogeneity analysis provides evidence that suggests (1) dynamic self-complementarity between first-year skills and the two-year productivity of the intervention program in terms of skill development and (2) dynamic cross-complementarity between first-year health and the two-year impact on skills.

This paper makes three primary contributions to the existing literature. First, we provide evidence on additivities between the impacts of a psychosocial stimulation and a health promotion component on child skill development and health outcomes of 1-year-olds in LMICs. Evidence on additive effects between psychosocial stimulation and health promotion programs during early childhood is scarce. To the best of our knowledge, no more than one earlier study documented this type of additivity for a two-year, postnatal home visitation program in LMICs. Specifically, Grantham-McGregor et al. (2020) used evidence from rural India to show that a psychosocial stimulation program can have a positive (one-year and two-year) impact on child skill development. Moreover, they report a positive one-

year impact of health education on child morbidity (i.e., the incidence of fever decreased by 11.8 percentage points). The positive impact on child morbidity, however, fades out during the second year of the intervention period (i.e., by age 3).

Second, this study proposes a novel reduced-form approach to explore dynamic complementarities between early capabilities and the two-year productivity of parenting training. Recently, a number of studies provided empirical evidence of dynamic complementarities by estimation of structural models of human capital formation (see, e.g., Attanasio, Cattan et al. 2020; Cunha and Heckman 2008). The identification of complex structural models demands a large sample size and a set of strong underlying assumptions (Chetty 2009). Structural models of human capital formation require, among others, assumptions about parental beliefs toward the expected returns on parental investment as well as about parental resource, time, and knowledge constraints (Del Boca et al. 2014). Moreover, no agreement exists on what the most fitting functional form is for human capital production functions: Cunha and Heckman (2008) assume a log-linear specification; Cunha et al. (2010) and Attanasio et al. (2017) assume a constant elasticity of substitution specification; and Attanasio, Cattan et al. (2020) and Attanasio, Meghir et al. (2020) assume a Cobb-Douglas specification. The current study, in contrast, uses a reduced-form approach that exploits the experimental variation in a purposeful way to explore dynamics in early capability formation and inform optimal program design while relying on fewer assumptions.

Third, this study investigates dynamic complementarities across different capability domains. Specifically, we study dynamic self-complementarity and cross-complementarity between the stock of early capabilities in a specific domain (i.e., skills or health) and the productivity of investments in this capability domain or another capability domain later on, respectively. Dynamic cross-domain complementarity implies that integration of program components can lead to

dynamic synergies between the impacts of program components (Alderman et al. 2014), which can play an important role in tailoring intervention programs to the developmental stage of children (Emmers et al. 2022). Even though this study does not apply a causal identification strategy for the estimation of production function parameters, the reduced-form results indicate that higher stocks of early child health entail dynamic cross-domain effects for the child's skill formation during the two-year intervention period.

The remainder of this paper is structured as follows. Section I provides the experimental literature on early capability formation in LMICs. Sections II and III present the experimental design and data collection, respectively. In Sections IV and V, we describe our estimation strategy and report the findings for concurrent and dynamic treatment impacts, respectively. Section VI provides the implications of our findings for our understanding of the impacts of parenting training programs and concludes.

I. Early Capability Formation and Policy Experiments in LMICs

Considering that an estimated 43% of the children under age 5 in LMICs are at risk of not reaching their developmental potential due to environmental risk factors (e.g., a lack of psychosocial stimulation), intervention programs targeted at early capability formation can play a key role in avoiding large losses of human potential and in building the human capital that a large share of the labor force will require as nations transition into higher-skilled economies. A systematic review and meta-analysis shows that 45%, 46%, and 36% of the children under age 5 in rural study sites across China (N = 19,762) are delayed in their cognitive, language, or social-emotional development, respectively (Emmers et al. 2021).

Based on the assertion that parenting training holds promise for producing significant and lasting gains in ECD outcomes, a range of parenting training

programs in LMICs have been implemented and evaluated. First, evidence from a pioneering parenting training program targeted at stunted Jamaican infants in the 1980s showed that parenting training in combination with nutritional supplementation improved cognitive skill formation and child growth in the short run (Grantham-McGregor et al. 1991) and a range of human capital outcomes, including improved academic attainment and adult labor productivity, 20 years later (Gertler et al. 2014; Walker et al. 2021). Following the 1990 World Declaration on Education for All, the findings of lasting impacts on a wide range of human capital outcomes of this relatively small-scale, targeted program (initially) with long-term follow-up (later) incentivized policymakers and academics to experiment with similar parenting training programs in numerous LMICs. In the 1990s and early 2000s, most programs were targeted at disadvantaged subpopulations of children, for example, underweight children, as seen in Walker et al. (2004) and Gardner et al. (2005). Over the past 15 years, non-targeted parenting training programs that involve the whole population of children in a region (e.g., those who were stunted and wasted and those who were not) have been implemented and evaluated. These parenting programs focused mainly on psychosocial stimulation, child health, or a combination of the two (see Emmers et al. (2022) for a review and discussion of parenting training programs in LMICs).

The empirical literature provides evidence of the short-run effectiveness of parenting training programs. First, psychosocial stimulation programs consistently benefit cognitive skill development. Jeong et al. (2021) conducted a meta-analysis that determined that parenting training programs that focus on child psychosocial stimulation in LMICs have, on average, a positive impact of 0.32 SD on child cognitive skills. The Chinese intervention literature confirms that psychosocial stimulation programs in rural China also consistently have positive effects on early cognitive development, with an average impact size of 0.26 SD (Emmers et al. 2021). Second, evidence on the effectiveness of education interventions that target

child health is mixed. Education on child health in most, but not all, studies produces an impact on child health or growth outcomes (see, e.g., Yousafzai et al. 2014; Luo et al. 2017). The one established pattern in the international literature is that health education programs without nutrition supplementation are, generally, insufficient to halt stunting in regions with high rates of malnutrition (Grantham-McGregor et al. 2014; Pérez-Escamilla and Moran 2017). Third, evidence on the effectiveness of integrated psychosocial stimulation and health programs shows that individual effects of single-component programs can be maintained (see, e.g., Grantham-McGregor et al. 2020). The findings indicate that parenting programs can function as tools to improve early skills and health in Chinese regions where access to ECD services (e.g., family support services, specialized counseling) is limited (Heckman et al. 2020; Sylvia et al. 2021).

In the short run, parenting training programs can affect child capabilities in mainly three ways: by (a) having a contemporaneous direct impact on child capabilities; (b) having an indirect impact on capability development via an increase in parental investment; and (c) leading to a change in the parameters of the production function. The theoretical framework of capability formation that was introduced by Cunha and Heckman (2007) captures these mechanisms:[1]

$$\theta_{t+1} = f_t(\theta_t, I_t, X_t) \tag{1}$$

According to this capability production function, capabilities at time $t+1$ (θ_{t+1}) are a function of capabilities at time t (θ_t), parental investment (I_t), and background characteristics (X_t).

[1] For analytical convenience, f_t is assumed to be strictly increasing in I_t, strictly concave in I_t, and twice continuously differentiable in all of its arguments. Consequently, more investments produce more capabilities $\frac{\partial \theta_{t+1}}{\partial I_t} > 0$.

In addition to identifying impacts on skills and health outcomes, studies have sought evidence of the underlying mechanisms behind treatment impacts. Studies consistently find positive effects on intermediate outcomes (that may be associated with the positive development outcomes), such as parental investments in child psychosocial stimulation and parenting knowledge (see, e.g., Andrew et al. 2018; Sylvia et al. 2021). Evidence from study sites in rural China shows that parenting training can raise parental engagement in interactive caregiver-child activities, such as telling stories and singing songs to children, and increase parenting knowledge (Emmers et al. 2021). The average impact sizes of 0.39 SD and 0.20 SD on parental investment in child psychosocial stimulation and parenting knowledge, respectively, suggest that parental investments in the child's home environment and changes in parental knowledge are likely to be two of the mechanisms behind the measured gains in child development. This argument is supported by structural parameter estimates of Attanasio, Cattan et al. (2020) that indicate that changes in parental investment explain a large share of the changes in skill development.

In the longer run, the persistence and size of treatment impacts also is determined by the expression and evolution of capabilities over time. At its core, the technology of capability formation in Equation (1) is a stage-specific development process that features dynamic productivity effects and complementarities (Cunha and Heckman 2007). If we define θ_t as a capability vector that contains child skill ($\theta_{S,t}$) and health capabilities ($\theta_{H,t}$) at time *t*:

$$\theta_t = \left(\theta_{S,t}\,, \theta_{H,t}\right) \tag{2}$$

and I_t as an investment vector that contains parental investments in child skill ($I_{S,t}$) and health capabilities ($I_{H,t}$) at time *t*:

$$I_t = \left(I_{S,t}\,, I_{H,t}\right), \tag{3}$$

we can describe dynamic complementarity as follows.[2]

- **Dynamic self-complementarity** is observed if $\frac{\partial^2 \theta_{S,t+1}}{\partial \theta_{S,t} \partial I_{S,t}} > 0$ or $\frac{\partial^2 \theta_{H,t+1}}{\partial \theta_{H,t} \partial I_{H,t}} > 0$. This means that the stock of a capability in period *t* makes investments in this capability during period *t* more productive. This implies that investments in a capability in period *t-1* bolster the productivity of investments in this capability in period *t*.[3]
- **Dynamic cross-complementarity** is observed if $\frac{\partial^2 \theta_{S,t+1}}{\partial \theta_{H,t} \partial I_{S,t}} > 0$ or $\frac{\partial^2 \theta_{H,t+1}}{\partial \theta_{S,t} \partial I_{H,t}} > 0$. This means that the stock of a capability in period *t* makes investments in another capability during period *t* more productive. This implies that investments in a capability in period *t-1* bolster the productivity of investments in another capability in period *t*.

Although empirical evidence on self-productivity, cross-productivity, and overall dynamic complementarity in human capital formation has started to emerge, evidence on dynamic self-complementarity and cross-complementarity is lacking. More specifically, a number of studies in LMICs, including the work of Attanasio et al. (2017), Attanasio, Cattan et al. (2020), and Attanasio, Meghir et al. (2020), find evidence of dynamic self-productivity and cross-productivity effects in the process of human capital formation during the first five years of life. Attanasio et al. (2017) and Attanasio, Meghir et al. (2020) find that child health at age 1 is

[2] Dynamic self-productivity is observed if $\frac{\partial \theta_{S,t+1}}{\partial \theta_{S,t}} > 0$ or $\frac{\partial \theta_{H,t+1}}{\partial \theta_{H,t}} > 0$, which means that a change in the stock of a capability in period *t* augments the development of this capability in period *t+1*. Dynamic cross-productivity is observed if $\frac{\partial \theta_{S,t+1}}{\partial \theta_{H,t}} > 0$ or $\frac{\partial \theta_{H,t+1}}{\partial \theta_{S,t}} > 0$, which means that a change in the stock of a capability in period *t* augments the development of another capability in period *t+1*.

[3] Considering the specification of the production function in Equation (1) and the assumption that $\frac{\partial \theta_{t+1}}{\partial I_t} > 0$, we know that if $\frac{\partial^2 \theta_{t+1}}{\partial \theta_t \partial I_t} > 0$ holds, then $\frac{\partial^2 \theta_{t+1}}{\partial I_{t-1} \partial I_t} > 0$ holds as well.

important for the cognitive development of 5-year-olds in Ethiopia, India, and Peru. Attanasio, Cattan et al. (2020) show that the cognitive development of Colombian 1-year-olds fosters social-emotional skill development at age 2. These findings indicate that it may be optimal to start investing early in developmental domains, such as child health and cognitive development. Moreover, a number of studies provide evidence of dynamic complementarity between early and late investments, using structural model estimation (see, e.g., Attanasio, Cattan et al. 2020; Cunha and Heckman 2008). No study, however, has attempted to disentangle dynamic self-complementarity from cross-complementarity.

Evidence on the dynamics of human capital formation can provide important information on the optimal timing of intervention programs. Evidence of dynamic complementarity between capabilities at an early stage of childhood and the productivity of an intervention later on can indicate that it is optimal to initiate the intervention program before this early stage of childhood. The presence of dynamic complementarities indicates that it is difficult for children with poor early capabilities to catch up with better-developed children. Moreover, dynamic complementarity is found to increase with age (Heckman and Mosso 2014). Hence, if interventions are initiated sufficiently early, then more children will be equipped with the right capabilities at the right time to seize emerging opportunities (e.g., children may need to acquire basic physical skills to be able to move around and develop more advanced cognitive skills later on; Johnson and Jackson 2019).

II. Experimental Design

A. Sampling and Randomization

We designed the intervention as a cluster-randomized controlled trial in Yunnan and Hebei Provinces, China (see Figure 1 for the trial profile). We listed all poor, rural townships in the respective provinces to select two townships in consultation

with local authorities. Two townships, one in each province, with an equally sized infant population were selected such that our sample was stratified by province. To reduce the risk of contamination across treatment and control group, we randomized at the village level. We used a list of all registered births in the selected villages (N = 43) to select the sample households. We enrolled all children who were between 6 and 18 months old on September 15, 2015, together with their primary caregivers. On average, 10 children per village were enrolled. The resulting baseline sample contains 449 caregiver-child dyads.

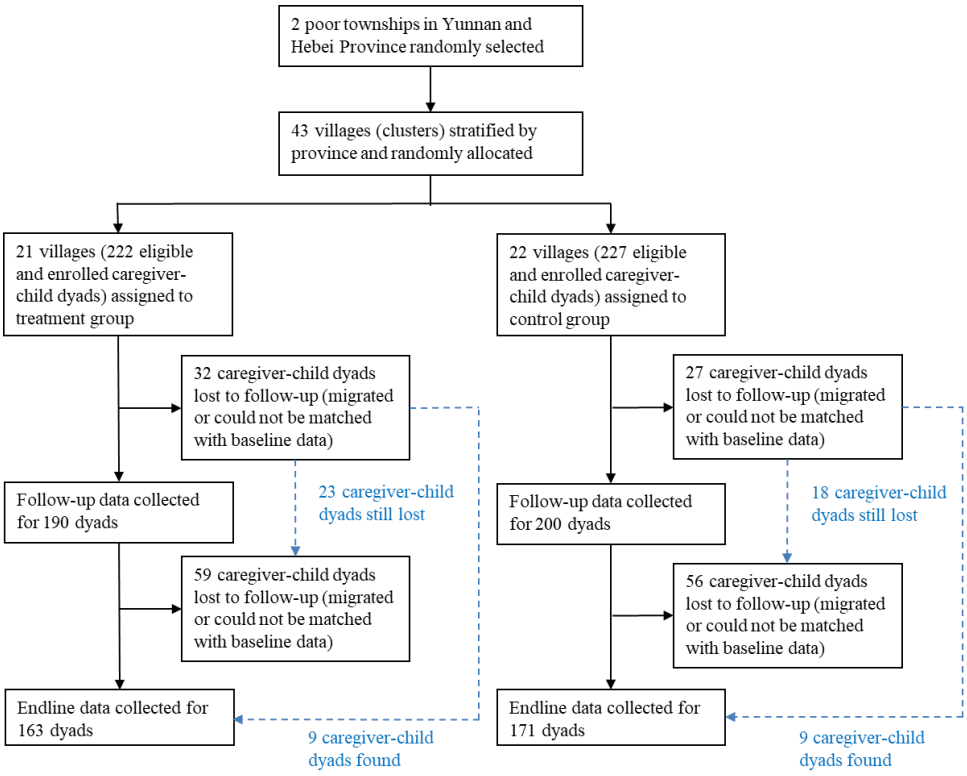


FIGURE 1. ITT TRIAL PROFILE

In September 2015, we used a computerized random number generator to randomly assign the villages to control and treatment groups within two strata (the sample township in Yunnan Province and the sample township in Hebei Province). Caregivers in the treatment group were unaware that they were involved in an experiment, and control group households were unaware of the intervention. Prior

to the start of the intervention, caregivers assigned to the intervention group were asked to provide oral informed consent for participation in the parenting intervention that was designed to last for two years, beginning in October 2015.

B. Intervention Program

All children and caregivers in treatment villages were invited to participate in a parenting training program on child psychosocial stimulation and health. During each of the biweekly home visits over a period of two years, community health workers (described below) trained caregivers in interactive caregiver-child activities and provided them with information on child health. All interactive caregiver-child activities were fully scripted in a stage-based ECD curriculum. This curriculum was based loosely on the curriculum developed for a pioneering parenting intervention in Kingston, Jamaica (Grantham-McGregor et al. 1991), and adapted to Chinese settings by local child development psychologists and ECD experts (Sylvia et al. 2021). Each activity focused on the development of one of four skill domains: cognition, language, motor, or social-emotional skills. Two new skill-building activities were introduced during each visit.

In addition, caregivers received information on child health through structured conversations about age-appropriate health topics at the end of each visit. These structured conversations were developed by experts from the Capital Institute of Pediatrics of the Chinese Academy of Medical Sciences. Health information covered issues that included age-appropriate feeding practices, immunization, hygiene, home safety, and sleeping behavior. Finally, the child was measured and weighed during each biweekly visit. In cases in which community health workers observed significantly deviating growth or development trajectories or signs of serious illnesses, they referred caregivers to a medic/doctor.

Home visits were delivered by community health workers who were hired and supervised by personnel from the local National Health Commission (NHC). To provide the community health workers with the necessary knowledge, skills, and tools, they received a copy of the ECD curriculum, an ECD toolkit (containing toys, picture books, and other counselling materials), and a manual with structured conversations on child health issues. In addition, all community health workers (*N* = 36, see Appendix Table A.1)[4] completed a one-week intensive training course, taught by three ECD experts, on theories of ECD, communication, coaching, and counseling. The training program combined five days of classroom-based instruction with two days of experiential learning in the field.

A carefully designed supervisory system was set up to monitor service quality. First, community health workers were requested to log into a mobile app at the start of each visit to track compliance. Second, county-level or township-level officials of the NHC randomly conducted unannounced observations of a home visit of each instructor every three months, provided feedback and, if needed, additional training. Third, NHC supervisors conducted phone interviews to invite caregivers to give feedback on the quality of home visits. All intervention households received phone calls on a revolving basis. Course content and reinforcement of the training sessions were adjusted according to caregiver feedback.

III. Data

A. Data Collection

Three rounds of data were collected on a yearly basis in 2015, 2016, and 2017 between late September and early October. The baseline sample included 449

[4] As presented in Appendix Table A.1, 63.9% of the trainers are female; 94.4% is married; 97.2% have children; and 41.7% hold a senior high school degree. None holds any type of tertiary education degree; 22.2% and 61.1% of the trainers work as government officials and farmers, respectively; and 83.3% are part of the village cadre.

caregiver child-dyads. Due to internal migration, 13% and 26% of the caregiver-child dyads had attrited by the time of midterm and endline data collection, respectively. As a consequence, follow-up data were collected for 390 and 334 caregiver-child dyads at midterm and endline, respectively. The caregivers of all children provided oral informed consent before baseline data collection.

Primary outcomes of interest were measures of three types of child capabilities: skill development, health, and physical growth. The Bayley-III test was administered to assess child skill development (Bayley 2006). The Bayley-III subscales for cognition, (receptive and expressive) language, (gross and fine) motor, and social-emotional skill development were previously translated into Mandarin Chinese and adapted to Chinese settings. The Mandarin Chinese versions of the Bayley-III test has been used in China and reported in studies of other research teams (Luo, Yue et al. 2017; Sylvia et al. 2021; Wang et al. 2019). Bayley-III tests were administered at home by trained enumerators who attended a one-week training course (including 2.5 days of field practice) ahead of data collection at baseline and both rounds of follow-up.

Given that Bayley-III raw scores increase by age and an official Chinese reference population distribution for normalization has not yet been established, we internally standardized Bayley-III raw scores using the observed distribution of scores in our sample to eliminate the age effect. Specifically, we first estimated age-conditional means and standard deviations (SD) for each age (month) group using a non-parametric regression method in line with Rubio-Codina et al. (2016). We then used these estimated statistics to compute age-adjusted internal *z*-scores for each subscale. We used a non-parametric standardization procedure, as this type of procedure is less sensitive to strong outliers and small numbers of observations within age groups than are parametric procedures.

In addition to measuring skill development, the research team evaluated health capabilities of children by measuring hemoglobin values, disease incidence, and

anthropometrics. Hemoglobin values were measured by HemoCue 201+ finger-prick blood tests. Finger-prick blood tests are the least invasive means and the standard public health method to determine hemoglobin values. All nursing staff who carried out HemoCue finger-prick blood testing were well trained, enrolled in nursing programs, and supervised by a certified nurse. Child disease incidence was measured by caregiver reports of the child's disease history. For example, the risk of diarrheal illness and respiratory tract infections were assessed based on caregiver-reported incidences of diarrheal-related illnesses and symptoms of respiratory tract infections (i.e., coughing or having a cold during the two weeks prior to the survey). Third, child height and weight were measured by trained enumerators. We used a scale to weigh children with minimal clothing. We did tared weighing of children below age 2 and children over age 2 who would not stand still. Height was assessed with a child measure mat or a stadiometer for children up to age 2 and children over age 2, respectively.

We further administered a comprehensive household questionnaire to each child's main caregiver to collect information on secondary outcomes, child characteristics, caregiver characteristics, and household characteristics. Secondary outcomes that might mediate intervention effects on primary outcomes in this study are measures of parenting knowledge and parental investment in child psychosocial stimulation and health (i.e., dietary diversity, nutrition supplementation, and home safety). In addition, we asked the respondent about child characteristics such as the child's sex, birth order, gestational age and weight at childbirth, and age (in months). We checked the child's sex and age on the child's official birth certificate. The survey team also collected information on the age, educational attainment, migration status, and household registration status of caregivers. Finally, we collected information on the household's recipient status of China's minimum living standard guarantee (*dibao*) program and access to certain types of assets (i.e., flush toilet, personal computer, and internet).

B. Baseline Characteristics, Balance, and Attrition

Table 1 shows summary statistics and tests for balance across control and treatment groups at baseline. Differences between treatment and control group in individual child and household characteristics are insignificant. A joint significance test in which we regress treatment status on all baseline characteristics confirms that the study arms are balanced (p = .722).

TABLE 1—BASELINE BALANCE TABLE

	Full sample (1)	Control (2)	Treatment (3)	(2) vs. (3), p-value
A) Child characteristics				
(1) Age (in months)	12.731	12.757	12.706	.760
	(0.180)	(0.289)	(0.219)	
(2) Male	0.465	0.432	0.500	.124
	(0.021)	(0.027)	(0.031)	
(3) Born prematurely	0.042	0.035	0.050	.566
	(0.012)	(0.015)	(0.019)	
(4) Firstborn	0.432	0.423	0.441	.619
	(0.024)	(0.034)	(0.035)	
(5) Cognition delay	0.487	0.447	0.527	.219
	(0.033)	(0.049)	(0.040)	
(6) Language delay	0.623	0.615	0.631	.981
	(0.041)	(0.042)	(0.071)	
(7) Motor delay	0.362	0.332	0.392	.341
	(0.030)	(0.047)	(0.036)	
(8) Social-emotional delay	0.533	0.527	0.541	.997
	(0.028)	(0.038)	(0.041)	
(9) Anemia (Hb < 110 g/L)	0.548	0.541	0.557	.508
	(0.030)	(0.041)	(0.046)	
(10) No. of times ill last month	0.909	0.960	0.856	.296
	(0.045)	(0.058)	(0.068)	
(11) Stunted (HAZ score < -2)	0.045	0.036	0.056	.387
	(0.010)	(0.012)	(0.016)	
(12) Wasted (WHZ score < -2)	0.023	0.022	0.023	.853
	(0.009)	(0.011)	(0.013)	
B) Household characteristics				
(1) Mother's age (in years)	27.900	27.890	27.910	.698
	(0.307)	(0.372)	(0.496)	
(2) Mother at home	0.889	0.894	0.883	.636
	(0.015)	(0.017)	(0.026)	
(3) Mother's education > 9 years	0.281	0.313	0.248	.228
	(0.026)	(0.037)	(0.033)	
(4) Household received social security support package	0.125	0.119	0.131	.783
	(0.016)	(0.023)	(0.021)	
No. of observations	449	227	222	

Notes: Mean and standard errors (in brackets) are reported. p-values account for clustering at the village level. Cognition, language, motor, and social-emotional delays are identified by a Bayley-III score that is more than 1 SD below the mean score of a healthy population.

The summary statistics on child characteristics in Panel A of Table 1 show that 46.5% of the children were male and that they were, on average, 12.7 months old

at baseline. In addition, 4.2% of the sample were prematurely born. Almost half (43.2%) of the sample were firstborns. Over 45% of the children had a cognition, language, or social-emotional delay, and 36.2% had a motor delay.[5] According to World Health Organization criteria, accounting for altitude adjustments, hemoglobin levels below 110 gram per liter (g/L) were counted as anemic. Measured this way, 45.2% of the children in the sample at baseline were anemic. Children had been ill, on average, 0.9 times over the month prior to the baseline. Using anthropometric data, we find that 2.3% or 4.5% of the children were wasted or stunted, respectively.

Panel B of Table 1 shows summary statistics for household characteristics. Mothers were, on average, 27.9 years old; 88.9% of the mothers were at home, taking care of their young child (who was in the sample), at baseline; 28.1% of the mothers had gone to school beyond junior high school (that is, they had more than 9 years of education); and 12.5% of the households were recipients of China's minimum living standard guarantee (*dibao*) program.

Further, we find no indication of attrition bias, where attrition is defined as having not been present to complete the household questionnaire at endline. According to the analysis, attrition is insignificantly correlated with treatment assignment ($p > 0.50$; see Appendix Table A.2). Summary statistics and tests for baseline balance for the sample of those did not attrite by the time of endline are reported in Appendix Table A.3. Differences between the control and treatment groups for individual and joint baseline characteristics remain insignificant after exclusion of the attriters. Hence, the evidence indicates that attrition was random.

[5] Delays in cognitive, language, motor, and social-emotional skill development were defined as a score of more than 1 SD below the mean of a healthily developed population. These cutoffs have been used in other research (Luo, Emmers et al. 2019; Wang et al. 2019).

IV. Average Treatment Effects on Primary and Secondary Outcomes

A. Estimation

Large and well-designed randomized experiments facilitate unbiased estimation of average treatment impacts (Angrist and Pischke 2010). We estimate treatment effects on the primary and secondary outcomes of interest. Provided that treatment assignment and attrition are random, as substantiated in Section III.B, comparing the means of outcomes between treatment arms results in unbiased estimates of the treatment effect. For estimation of intention-to-treat (ITT) effects of the intervention on primary and secondary outcomes, we use ordinary least squares (OLS) to estimate the following ANCOVA specification:[6]

$$Y_{ijt} = \alpha_0 + \alpha_1 T_j + \alpha_2 Y_{ij0} + \pi_s + \varepsilon_{ijt}, \tag{4}$$

where Y_{ijt} is a primary or secondary outcome measure for child *i* in village *j* at the end of intervention year *t*, where *t* takes the value 1 or 2; T_j is a dummy variable that indicates treatment assignment of village *j*; Y_{ij0} is the outcome measure for child *i* at baseline, and π_s is a set of province fixed effects. For estimation of ITT effects on skill development scores, we include Bayley tester fixed effects. We adjust standard errors for clustering at the village level, using the cluster-correlated Huber-White estimator.

Because we estimate treatment effects on a diverse set of primary and secondary outcomes, we need to account for the risk of multiple inference (i.e., the increased likelihood of false discoveries when examining multiple outcomes using fixed *p*-values for each hypothesis). We account for multiple hypothesis testing in two

[6] In addition to ITT effects, we estimate dose-response relationships to investigate the marginal impact per completed home visit on primary and secondary outcomes (see Appendix B). Appendix B provides the results for the analysis of the drivers and impacts of compliance.

ways. First, we aggregate outcome measures within each domain into a summary index and estimate ITT effects on these summary indices. Following the methodology of Fitzsimons et al. (2016), each index is computed as a weighted mean of the standardized values of the outcome variables (with outcome variables redefined such that higher values translate into more desirable outcomes). This summary index is an efficient generalized least squares (GLS) estimator. The GLS weights are calculated to maximize efficiency by giving less weight to outcomes that are highly correlated with each other, while outcomes that are uncorrelated, and, thus, contain new information, receive more weight. To obtain meaningful ITT estimates, we internally standardize the summary indices within age (month) group. Second, using the methodology of Romano and Wolf (2016), we adjust *p*-values for multiple hypotheses using a step-down procedure, which controls for the familywise error rate in each family of outcomes. The outcome variables are grouped by the family of outcomes in Appendix Tables A.5 and A.7–A.13. We conduct the statistical analysis using Stata version 17.0. We use the RWOLF package to implement the Romano-Wolf stepdown procedure (Clarke 2020).

Moreover, as child capabilities and parental investment are inherently unobservable (Cunha and Heckman 2008), the measures described in Section III.A are, to a certain extent, error-ridden indicators of the underlying latent variables of child capabilities and parental investment. Using any one set of these measures instead of the latent variables could lead to severely biased results, whether or not the model is linear (Attanasio, Cattan et al. 2020). We use the summary indices as measures of the latent variables.

B. Results

Figure 2 (Rows 1 to 3) displays ITT effects on the summary indices for skill development, health, and physical growth.[7] After one year, the intervention improves skill development and health capabilities by 0.18 SD and 0.25 SD, respectively. By the end of the second year, the intervention effect on skill development increases to 0.33 SD. In the case of the health index, however, the impact fades out. The intervention program has no impact on physical growth in either the first or second year. ITT effects on the individual outcome variables that were used to construct the summary indices of skill development, health, or physical growth are reported in Appendix Tables A.5, A.7, and A.8.[8]

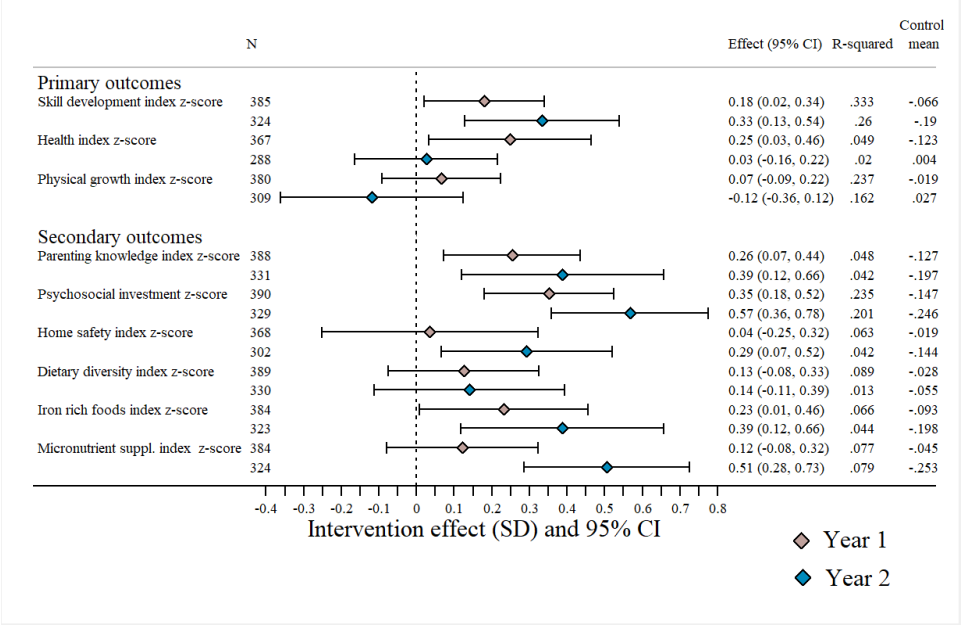


FIGURE 2. ITT EFFECTS ON SUMMARY INDICES

[7] As a robustness check, we estimate ITT effects on summary indices, while controlling for a list of additional control variables (see Appendix Table A.4). The optimal set of additional controls (i.e., the set that minimizes the risks of omitted variable bias and overfitting) was selected using the post-double-selection methodology of Belloni et al. (2016) implemented with Stata's pdslasso package. The pdslasso package estimates two lasso regressions to select the union of controls that significantly affect the outcome of interest as well as treatment assignment. Controlling for additional control variables has a marginal impact on the estimated ITT effects (see Appendix Table A.4).

[8] Appendix Table A.5 displays ITT effects on standardized Bayley-III z-scores. ITT effects on raw Bayley-III aggregate item scores are reported in Appendix Table A.6.

Notes: Effects on primary and secondary outcomes are estimated intervention effects that control for baseline scores and province fixed effects. To estimate ITT effects on skill development, we also controlled for tester fixed effects. Standard errors in are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level.

In regard to the impacts on secondary outcomes (see Figure 2, Rows 4 to 9), the results indicate that the parenting training program also affects a number of secondary outcomes. According to the analysis, parenting knowledge increases by 0.25 SD and 0.39 SD after one year and two years, respectively. Parental investment in psychosocial stimulation and iron-rich foods also increases by 0.35 SD and 0.23 SD, respectively, after one year. After two years, the magnitude of these impacts increases to 0.57 SD and 0.39 SD, respectively. The home safety index increases by 0.29 SD after two years of intervention. ITT effects on the individual outcome variables that were used to construct the summary indices for the secondary outcomes are reported in Appendix Tables A.9–A.13.[9]

V. Dynamic Effects

A. Estimation

This study investigates how the capability formation of individuals during different stages of childhood affects the productivity of policy interventions during the remainder of the intervention period. In line with the work of Aizer and Cunha (2012), we assume that the interaction effect between the initial stock of capabilities and randomized treatment assignment (i.e., a source of exogenous variation in parental investment) captures the dynamic complementarity between early capabilities and the productivity of later investments in the production of capabilities. Considering that we have three data waves, we are able to estimate how two-year treatment impacts differ across children with favorable or

[9] We also tested for intervention effects on nursery school enrollment and nursery school enrollment age (in months). The results presented in Appendix Table A.14 show no significant impact of the treatment on enrollment. We do find, however, that children from Hebei Province who were assigned to the treatment group were, on average, enrolled in nursery school 1.2 months earlier than were children who were assigned to the control group.

unfavorable baseline outcomes (i.e., at age 1) and across children with favorable or unfavorable (impacts on) outcomes at the end of the first year of the intervention (i.e., at age 2). Taking into account that randomized experiments allow for unbiased estimation of treatment effects by subgroup (Bitler et al. 2017), we use a subgroup heterogeneity analysis to explore how differences in initial capabilities induce heterogeneity in the productivity of a two-year parenting intervention program.

As a general rule, subgroups for impact heterogeneity analysis must be created based on characteristics that are immutable or observed before randomization (Bitler et al. 2017). First, we define subgroups based on baseline capabilities. Second, we use in-sample information on relationships between a comprehensive set of baseline covariates (see Appendix Table A.15) and child capabilities (i.e., skill development and health) at the end of Year 1 to predict first-year outcomes for all experimental units. Specifically, we investigate relationships between baseline covariates and (1) first-year outcomes in the control group to predict potential outcomes in the absence of treatment and (2) first-year outcomes in the treatment group to predict potential outcomes in the presence of treatment.

It is important to note that, traditionally, control group data are used for the prediction of outcomes for endogenous stratification (Abadie et al. 2018). When using treatment group data for prediction as well, we are able to explore how the predisposition to acquire new capabilities during the first intervention year can alter the productivity of later parental investments. We assume that using the distribution of outcomes in the treatment group for prediction does not cause endogeneity bias as long as the predicted outcomes are uncorrelated with treatment assignment. To strengthen this assumption, we (1) do not use the dummy of treatment assignment as a predictor for first-year outcomes; (2) verify that predictors are not correlated with treatment assignment (see Tables 1 and A.3); (3) use the same prediction sample to predict outcomes for the full sample; and (4) test for balancedness of the predicted outcomes across the control and treatment groups.

Because predictions of outcomes based on classical regression models may lead to substantially biased estimates of treatment effects in endogenously stratified subgroups due to overfitting (Abadie et al. 2018), we compare prediction performance of OLS regressions, penalized (i.e., lasso and post-lasso OLS) regressions,[10] and random forest machine learning algorithms[11] based on cross-validated MSPE.[12] We find that the average MSPE of the random forest estimator is lowest for skill development and health scores (see Appendix Table A.16). Hence, this study uses the random forest algorithm for the first-step prediction problem of the endogenous stratification estimator.

In addition to the prediction of skill and health outcomes at the end of the first year, we predict first-year impacts on the outcomes using the causal random forest algorithm that was developed by Athey et al. (2019). This algorithm predicts impacts as the predicted conditional average treatment effect, given baseline covariates. Appendix Figure A.1 provides a plot of the kernel density estimates of the distribution of the (causal) random forest-based predictions of outcomes and impacts. Appendix Figure A.2 presents a plot of the kernel density estimates by treatment assignment. The equality between the treatment and control group, with regard to the mean and distribution of the predicted outcomes and impacts, cannot

[10] The (square-root) lasso estimator minimizes the (root) mean squared error subject to a penalty on the absolute size of the coefficient estimates. Ahrens et al. (2020) explain these penalized regression methods in detail. We use the rlasso command of the Stata lassopack package that implements the theory-driven penalization methodology of Belloni et al. (2016) for (square-root) lasso regressions. The rlasso command allows us to account for clustering of errors at the village level. The *lasso2* and *cvlasso* commands of the lassopack package, in contrast, do not allow accounting for clustered errors.

[11] We use the Stata MLRtime package of Huntington-Klein (2020) to run the regression forest and causal forest algorithms of the generalized random forest (grf) package in R (Tibshirani et al. 2020). These commands allow accounting for clustering at the village level. We use honest splitting, and parameters are tuned by cross-validation.

[12] We conduct a *K*-fold cross-validation analysis (Kuhn and Johnson 2013), whereby we first split the data in K approximately equally sized groups, or "folds." Each fold k contains n_k observations for $k = 1, \dots, K$. $K_k$is the set of observations in each data partition. In the k^{th} step, the k^{th} fold is treated as the validation data set, and the remaining $K - 1$ folds constitute the training data set. The resulting estimate, which is based on all of the data, except for the observations in fold k, is $\hat{\beta}_{-k}$. The procedure is repeated for each fold such that every data partition is used for validation once. In line with Ahrens et al. (2020), we compute the *K*-fold cross-validation MSPE (CV_K^{MSPE}) for each prediction algorithm as follows:

$$CV_K^{MSPE} = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{n_k}\sum_{i\in K_k}\left(y_i - x_i'\hat{\beta}_{-k}\right)^2\right)$$

be rejected at the 10% level. This indicates that the predicted outcomes and impacts are unlikely to be correlated with treatment assignment.

In the next step, we use these predictions to define subgroup indicators. We define subgroup indicators for children with (impacts on) capabilities above or below the median for a traditional subgroup heterogeneity analysis. We estimate the following OLS regression specification:

$$(5) \quad Y_{ij2} = \alpha_0 + \alpha_1 T_j G_{yijt}^{HIGH} + \alpha_2 T_j G_{yijt}^{LOW} + \alpha_3 G_{yijt}^{HIGH} + \alpha_4 Y_{ij0} + \pi_s + \varepsilon_{ijt},$$

where Y_{ij2} presents a vector of capabilities of child *i* in village *j* at the end of the second year. This vector includes skills (S_{ij2}) and health capabilities (H_{ij2}). $T_j G_{yijt}^{HIGH}$ and $T_j G_{yijt}^{LOW}$ are two dummy variables that indicate that child *i* is living in treatment group village *j*, and the (impact on) capability *y* (i.e., skill (S) or health (H)) of child *i* in village *j* at baseline ($t = 0$) or at the end of the first year ($t = 1$) is above or below the median, respectively. G_{yijt}^{HIGH} is a dummy variable that takes the value of 1 if the (impact on) capability *y* of child *i* in village *j* is above the median, and the value of 0 otherwise. Further, we include controls for the capability score at baseline (Y_{ij0}), province fixed effects (π_s), and for the estimation of treatment impacts on child skills, tester fixed effects.

Under the assumption that treatment assignment induces an exogenous shock in the level of investment in child development, the difference ($\alpha_1 - \alpha_2$) in regression specification (5) captures the presence of dynamic complementarity between early capabilities and later investments in the production of capabilities. If dynamic complementarity is present, then we expect the difference ($\alpha_1 - \alpha_2$) to be significant and positive. If the dependent variable (Y_{ij2}) matches the capability measure *y* used to define subgroup dummy G_{yij1}^{HIGH}, then ($\alpha_1 - \alpha_2$) captures dynamic self-complementarity. If the dependent variable (Y_{ij2}) does not match the

capability measure *y* used to define subgroup dummy G_{yij1}^{HIGH}, then we assume that $(\alpha_1 - \alpha_2)$ captures dynamic cross-complementarity.

Finally, we acknowledge that the proposed reduced-form approach is not a causal identification strategy. The estimation strategy relies on the assumption that treatment assignment induces an exogenous shock in the level of parental investment and that the size of this shock is unrelated to children's initial capabilities. If treatment impacts on parental investments are larger for children with higher initial capabilities, then $(\alpha_1 - \alpha_2)$ may pick up the impact of reinforcing parental investments and, as a consequence, overestimate the dynamic complementarity in capability formation. If treatment impacts on parental investment are larger for children with lower initial capabilities, then $(\alpha_1 - \alpha_2)$ may be contaminated by compensatory investments and, as a consequence, underestimate the dynamic complementarity. In order to test this assumption, we assess subgroup heterogeneity in the treatment impacts on parental investments between children with high and low initial capabilities. Moreover, previous studies have shown that endogenous stratification bias can lead to a downward and upward bias of the treatment effect estimators for the higher (α_1) and lower (α_2) predicted outcome groups, respectively (Abadie et al. 2018). Therefore, we expect that, if overfitting bias occurs, this bias is more likely to lead to an underestimation than to an overestimation of the dynamic complementarity $(\alpha_1 - \alpha_2)$.

B. Results

First, we investigate subgroup heterogeneity between children with high and low baseline capabilities (i.e., skill development or health). The results in Columns 1 and 4 of Table 2 show that two-year ITT effects on skill development do not differ significantly between children with low and high baseline capabilities. We infer that we find no evidence of dynamic complementarity between baseline skill

development or health and the two-year productivity of the parenting training program in terms of skill development. The data presentation in the left panel of Appendix Figure A.3 also confirms that children with low baseline skills are able to benefit from the parenting intervention and to catch up with their more advantaged peers during the two-year intervention period.

TABLE 2—HETEROGENEOUS ITT EFFECTS ON SKILL DEVELOPMENT AT END OF YEAR 2 BY BASELINE OR PREDICTED SKILL DEVELOPMENT OR HEALTH AT END OF YEAR 1

	Dependent variable: Skill development index *z*-score at end of year 2					
	Subgroups: Low vs. high skill development			Subgroups: Low vs. high health		
	At baseline	Year 1 prediction based on control group	Year 1 prediction based on treatment group	At baseline	Year 1 prediction based on control group	Year 1 prediction based on treatment group
	[1]	[2]	[3]	[4]	[5]	[6]
Estimated ITT effect						
Low-outcome group	0.326**	0.272*	0.143	0.217	0.128	0.245*
	(0.135)	(0.144)	(0.116)	(0.141)	(0.13)	(0.128)
High-outcome group	0.336**	0.295*	0.462***	0.437***	0.554***	0.414***
	(0.152)	(0.153)	(0.155)	(0.146)	(0.142)	(0.135)
Skill development index *z*-score at baseline	0.280***	0.203***	0.197***	0.239***	0.261***	0.237***
	(0.082)	(0.049)	(0.051)	(0.051)	(0.053)	(0.05)
R^2	0.261	0.295	0.283	0.258	0.272	0.262
Observations	324	324	324	314	324	324
p-value of Wald test for homogeneity of estimated ITT effects	.960	.914	.094	.286	.021	.319

Note. The treatment effects on skill development in Columns [1]–[6] are regression coefficients on the treatment variable by predicted subgroup in a linear regression that includes controls for the predicted outcome group, baseline skill development, and province and tester fixed effects. Skill development and health outcomes at the end of the first intervention year are predicted using a random forest algorithm. Standard errors (in brackets) are computed with the cluster-correlated Huber-White estimator with clustering at the village level.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Two-year ITT impacts on child health are significantly larger for children with good baseline child health (with an ITT estimate of 0.29 SD; see Row 2 of column 4 of Appendix Table A.17) than for children with low baseline health (with an ITT estimate of -0.23 SD; see Row 1). This indicates that health outcomes formed before the start of the intervention (i.e., before age 1) may raise the productivity of the health promotion program component. The data presentation in the right panel

of Appendix Figure A.3 shows that, in disregard of the initial catch-up during the first intervention year, the health outcomes of children with low baseline health are lagging behind at the end of the second year of the intervention.

We then investigate heterogeneity in two-year ITT effects between subgroups of children with high and low predicted first-year capabilities. Column 3 of Table 2 shows that the two-year ITT effect on skill development is smaller for children with low predicted first-year skill development after treatment (with an ITT estimate of 0.14 SD; see Row 1) than for the ones with high predicted first-year skills after treatment (with an ITT estimate of 0.46 SD; see Row 2).[13] We interpret this difference in two-year treatment impacts as evidence of dynamic self-complementarity in the process of skill development. The difference in treatment impacts is significant at the 10% level when we use treatment group data for prediction (see Column 3) but not when we use control group data (see Column 2). Further, the results in Columns 5 and 6 of Table 2 show that two-year ITT effects on skill development are smaller for children with low predicted first-year health (with ITT estimates that range from 0.13 to 0.24 SD; see Row 1) than for the ones with high predicted first-year health (with ITT estimates that range from 0.41 to 0.55; see Row 2). This difference provides evidence of dynamic cross-complementarity between health capabilities and skill development. The difference in treatment impacts is significant at the 5% level when we use control group data for prediction (see Column 5), but not significant at the 10% level when we use treatment group data (see Column 6).[14]

[13] We conduct a parallel analysis of the heterogeneity in ITT effects on child health between subgroups with high and low predicted first-year outcomes (see Columns 2 and 3, 5 and 6, and 8 and 9 of Appendix Table A.17). The impacts on child health are systematically smaller for children with poor predicted skill development or health at the end of the first intervention year (with ITT estimates that range from -0.14 to -0.01 SD; see Row 1) than for children with good predicted year 1 outcomes (with ITT estimates that range from 0.07 to 0.17 SD; see Row 2). In line with the results presented in Table 2, however, we find no evidence of significant two-year impacts on child health.

[14] Appendix Table A.18 presents heterogeneity in ITT effects on skill development and health between subgroups of children with high or low predicted first-year impacts. We find no evidence of significant differences in two-year impacts on skill development or health across the predicted first-year impact groups (see Columns 1–4; $p > 0.10$).

As seen in Figure 3, we combine initial skill development and health outcomes to define two subgroups: children with low and high initial capabilities, for which initial capabilities comprise both skills and health. We find that the treatment has a small two-year impact on children with low initial skills and health (with ITT estimates that range from 0.09 to 0.23 SD; see left panel). For children with high initial skills and health, in contrast, the intervention has, on average, a two-year impact of more than 0.45 SD (with ITT estimates that range from 0.45 to 0.61 SD). The difference in treatment impacts is significant at the 10% significance level when we use treatment group data for random forest prediction of Year 1 outcomes. We find no significant evidence of treatment impacts on child health or heterogeneity in the treatment impacts (see the right panel of Figure 3).

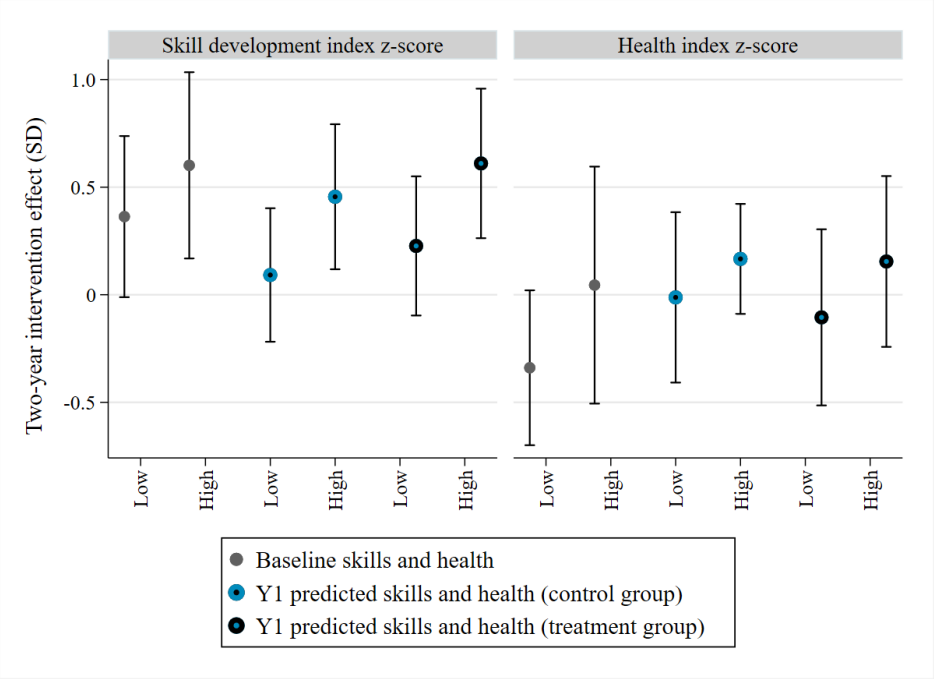


FIGURE 3. HETEROGENEITY IN TWO-YEAR ITT EFFECTS ON CHILD CAPABILITIES BY BASELINE OR PREDICTED SKILL DEVELOPMENT AND HEALTH AT END OF YEAR 1

Note. The treatment effects on skill development and health in Columns [1] – [12] are regression coefficients on the treatment variable by predicted subgroup in a linear regression that includes controls for the predicted outcome group, baseline skill development, and province fixed effects. To estimate treatment impacts on skill development, we include tester fixed effects. Standard errors are computed with the cluster-correlated Huber-White estimator with clustering at the village level. The *p*-value of the difference in treatment impacts on skill

development of children with low and high predicted year-1 skills in Columns [5] and [6] is significant at the 10% level (p = .075). The remaining differences are not significant at the 10% level.

Finally, Appendix Table A.19 presents the results when we investigate subgroup heterogeneity in the treatment impacts on parental investment in child psychosocial stimulation. The results provide limited evidence of heterogeneity. This indicates that treatment assignment induces an exogenous shock to the level of parental investment in child development that is unrelated to children's initial capabilities. If anything, treatment impacts on caregivers of children with low initial capabilities are higher than the impacts on their counterparts. The treatment impact on caregivers of children with low baseline health is higher (with an ITT estimate of 0.76 SD; see Column 4, Row 1) than for caregivers with children with high baseline health (with an ITT estimate of 0.35 SD; see Column 4, Row 2). This indicates that the results in Table 2 and Figure 3 are more likely to underestimate than to overestimate the role of dynamic complementarities in capability formation.

VI. Discussion and Conclusion

The overall goal of this study is to investigate the mechanisms through which a two-year, home-based parenting training program drives the capability formation of young children (1 year old at baseline) in rural areas of China. Using data from a randomized controlled trial that includes three data waves (baseline, one year after the start of the intervention, and endline after two years), we are able to use empirical methods to measure short-run (one-year) and longer-run (two-year) impacts, identify mediating effects, and search for dynamic (self- and cross-) complementarity. Our results show that an integrated parenting training program can be effectively delivered by community health workers with limited educational background and average socioeconomic status. The detected one-year and second-year ITT effects on skill development of 0.18 SD and 0.33 SD, respectively, are in line with the average treatment impacts of 0.26 SD and 0.32 SD, respectively, on skill development reported in other parenting experiments focusing on

psychosocial stimulation of infants and toddlers in rural China (Emmers et al. 2021) and other LMICs (Jeong et al. 2021).

The integrated parenting curriculum with a psychosocial stimulation and a health component in this experiment allows us to test for the presence of synergies between the two components. We find that, in addition to gains in skill development, this type of intervention program can lead to an improvement in child health of 0.25 SD during the first year of the intervention. This finding has important implications for policymakers interested in the design of cost-effective and scalable interventions. To the extent in which benefits of gains in capabilities exceed incremental costs of adding a program component, integrating program components can improve the cost-effectiveness of programs (Engle et al. 2007).

The impact on child health, however, is shown to fade out during the second year of the intervention, which may be caused by a natural adjustment in child health (for those in the control group) that occurs by the time the child turns 3. This type of interpretation would be consistent with the findings of Luo et al. (2017), who observe that child health outcomes (i.e., hemoglobin values and anemia rates) naturally improve between the ages of 18 and 30 months of age in rural China. Specifically, by monitoring the control group in their study, the authors find a steady increase (of 9 g/L) in hemoglobin levels and a steady decrease (of 25%) in anemia prevalence in absence of intervention. Consequently, Luo et al. find that the initial positive impact of a six-month micronutrient supplementation program on hemoglobin levels of 1-year-olds disappears (relative to the control group) by the time they turn 2.5 years old. For our sample, we observe, that at age 2, 6% and 12% of the children in the treatment and control groups, respectively, had diarrhea over the past two weeks. By age 3, 7% and 5% of the children in the treatment and control groups, respectively, had diarrhea during the past weeks.

Our results also shed light on the mechanisms through which the program affects child skill and health capabilities. The program has significant impacts on parenting

knowledge and a set of parental investments, such as parental investment in psychosocial stimulation, home safety, the provision of iron-rich foods in the diets of their children, and the use of micronutrient supplements. Impacts on secondary outcomes are more significant and larger at the end of the second year than at the end of the first year. This indicates that the parents are willing to continue to learn about and invest in their child's development and health during the second year of the intervention program.

In addition, the results of this study show that two-year impacts on skill development are significantly larger for children with higher predicted first-year skill development and/or health. These findings provide suggestive evidence of dynamic synergies between investments in different types of capabilities at different stages of childhood. On the one hand, two-year productivity of the intervention program in terms of skill development is higher for children who are predicted to have higher skills at the end of the first year of the intervention in the presence of treatment. Hence, this indicates the presence of dynamic self-complementarity in the skill-building process at age 2. On the other hand, we find evidence to indicate dynamic cross-complementarity between higher first-year health outcomes and two-year treatment impacts on skill capabilities. The finding that the initial stock of skills raises the productivity of later skill investment is in line with findings of dynamic self-complementarity in the literature (Attanasio, Catten et al. 2020; Cunha et al. 2010). To the best of our knowledge, this study provides the first empirical evidence of dynamic cross-complementarity between health produced at one stage of childhood and the productivity of investments in skill development during a subsequent stage.

The evidence of dynamic effects between initial capabilities and the productivity of the intervention program has implications for the optimal timing of parenting interventions. First, the finding that two-year ITT effects on skill development do not differ significantly between children with low and high baseline capabilities

suggests that 1-year-olds in rural China have the basic capabilities required for further skill building. Children with low baseline skills are able to catch up with more advantaged peers during the two-year intervention period. In combination with the findings of dynamic complementarities between predicted capabilities at age 2 and two-year program effectiveness, we conclude that it may be optimal to implement a psychosocial stimulation program at age 1 or during the second year before the child reaches the age of 2. Second, the finding that health outcomes formed before the start of the intervention (i.e., before age 1) can raise the productivity of the health promotion component indicates that children with poor health at age 1 may lack foundational building blocks for long-term health. It may be optimal to initiate health promotion programs during the first months of life or even before childbirth.

Although we believe that our research has produced a robust set of results with a number of important findings, we also recognize that the study includes a number of limitations. First, the study took place in two rural townships in China, and the results may be different in other contexts. Nevertheless, the sample appears to be fairly representative of caregiver-child dyads in remote villages in rural China. Second, we estimate ITT effects only at the midterm and endline periods. Longer-term follow-up of sample children is necessary to determine whether observed gains can persist over a longer period of time after the end of the intervention program. Third, we acknowledge that using the reduced-form approach to explore dynamic complementarity in capability formation is not a causal identification strategy. As explained in Section V, the reported results are more likely to underestimate than to overestimate the role of dynamic complementarity due to endogenous stratification bias. Finally, we acknowledge that, due to the fact that psychosocial stimulation and health promotion training was delivered to all treated subjects, we are unable to fully disentangle the impacts of the two intervention components. Future studies need to study dynamic complementarities between

program impacts based on evidence from multi-arm, multi-stage randomized controlled trials with re-randomized treatment assignment at each stage.

REFERENCES

Abadie, Alberto, Matthew M. Chingos, and Martin R. West. 2018. "Endogenous Stratification in Randomized Experiments." *Review of Economics and Statistics* 100 (4): 567–580. https://doi.org/10.1162/rest_a_00732.

Ahrens, Achim, Christian B. Hansen, and Mark E. Schaffer. 2020. "Lassopack: Model Selection and Prediction with Regularized Regression in Stata." *The Stata Journal: Promoting Communications on Statistics and Stata* 20 (1): 176–235. https://doi.org/10.1177/1536867X20909697.

Aizer, Anna, and Flávio Cunha. 2012. "The Production of Human Capital: Endowments, Investments and Fertility." Working Paper 18429, National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/W18429.

Alderman, Harold, Jere R. Behrman, Sally Grantham-McGregor, Florencia Lopez-Boo, and Sergio Urzua. 2014. "Economic Perspectives on Integrating Early Child Stimulation with Nutritional Interventions." *Annals of the New York Academy of Sciences* 1308 (1): 129–138. https://doi.org/10.1111/nyas.12331.

Almond, Douglas, and Janet Currie. 2011. *Human Capital Development before Age Five*. *Handbook of Labor Economics*. Volume 4. North Holland: Elsevier.

Andrew, Alison, Orazio Attanasio, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina. 2018. "Impacts 2 Years after a Scalable Early Childhood Development Intervention to Increase Psychosocial Stimulation in the Home: A Follow-up of a Cluster Randomised Controlled Trial in Colombia." *PLoS Medicine* 15 (4). https://doi.org/10.1371/journal.pmed.1002556.

Angrist, Joshua D., and Jörn Steffen Pischke. 2010. "The Credibility Revolution in

Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30. https://doi.org/10.1257/jep.24.2.3.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47 (2): 1179–1203. https://doi.org/10.1214/18-AOS1709.

Attanasio, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. 2020. "Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia." *American Economic Review* 110 (1): 48–85. https://doi.org/10.1257/aer.20150183.

Attanasio, Orazio, Costas Meghir, and Emily Nix. 2020. "Human Capital Development and Parental Investment in India." *The Review of Economic Studies* 87 (6): 2511–2541. https://doi.org/10.1093/restud/rdaa026.

Attanasio, Orazio, Costas Meghir, Emily Nix, and Francesca Salvati. 2017. "Human Capital Growth and Poverty: Evidence from Ethiopia and Peru." *Review of Economic Dynamics* 25: 234–259. https://doi.org/10.1016/j.red.2017.02.002.

Bayley, N. 2006. *Bayley Scales of Infant and Toddler Development–Third Edition*. San Antonio, TX: Harcourt Assessment.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. 2016. "Inference in High-Dimensional Panel Models With an Application to Gun Control." *Journal of Business and Economics Statistics* 34 (4): 590–605. https://doi.org/10.1080/07350015.2015.1102733.

Bitler, Marianne P, Jonah B Gelbach, and Hilary W Hoynes. 2017. "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment." *The Review of Economics and Statistics* 99 (4): 683–697. https://doi.org/10.1162/REST_a_00662.

Black, Maureen M., Susan P. Walker, Lia C.H. Fernald, Christopher T. Andersen, Ann M. DiGirolamo, Chunling Lu, Dana C. McCoy, et al. 2017. "Early

Childhood Development Coming of Age: Science through the Life Course." *The Lancet* 389 (10064): 77–90. https://doi.org/10.1016/S0140-6736(16)31389-7.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods." *Annual Review of Economics* 1 (1): 451–88. https://doi.org/10.1146/annurev.economics.050708.142910.

Clarke, Damian. 2020. "RWOLF: Stata Module to Calculate Romano-Wolf Stepdown p-Values for Multiple Hypothesis Testing." *Statistical Software Components*. https://EconPapers.repec.org/RePEc:boc:bocode:s458970.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97 (2): 31–47. https://doi.org/10.1257/aer.97.2.31.

Cunha, Flavio, and James J. Heckman. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Source: The Journal of Human Resources* 43 (4): 738–782. doi: 10.3368/jhr.43.4.738.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931. https://doi.org/10.3982/ecta6551.

Del Boca, Daniela, Christpher Flinn, and Matthew Wiswall. 2014. "Household Choices and Child Development." *The Review of Economic Studies* 81 (1): 137–185. https://doi.org/10.1093/restud/rdt026.

Emmers, Dorien, Juan Carlos Caro, Scott Rozelle, and Sean Sylvia. "Early Parenting Interventions to Foster Human Capital in Developing Countries." *Annual Review of Resource Economics* 14. Published electronically June 13, 2022.

Emmers, Dorien, Qi Jiang, Hao Xue, Yue Zhang, Yunting Zhang, Yingxue Zhao, Bin Liu, et al. 2021. "Early Childhood Development and Parental Training Interventions in Rural China: A Systematic Review and Meta-Analysis." *BMJ*

Global Health 6 (8): e005578. https://doi.org/10.1136/BMJGH-2021-005578.

Engle, Patrice L., Maureen M. Black, Jere R. Behrman, Meena Cabral de Mello, Paul J. Gertler, Lydia Kapiriri, Reynaldo Martorell, and Mary Eming Young. 2007. "Strategies to Avoid the Loss of Developmental Potential in More than 200 Million Children in the Developing World." *Lancet* 369 (9557): 229–242. https://doi.org/10.1016/S0140-6736(07)60112-3.

Fitzsimons, Emla, Bansi Malde, Alice Mesnard, and Marcos Vera-Hernández. 2016. "Nutrition, Information and Household Behavior: Experimental Evidence from Malawi." *Journal of Development Economics* 122: 113–126. https://doi.org/10.1016/j.jdeveco.2016.05.002.

Gardner, Julie M. Meeks, Christine A. Powell, Helen Baker-Henningham, Susan P. Walker, Tim J. Cole, and Sally M. Grantham-McGregor. 2005. "Zinc Supplementation and Psychosocial Stimulation: Effects on the Development of Undernourished Jamaican Children." *American Journal of Clinical Nutrition* 82 (2): 399–405. https://doi.org/10.1093/ajcn/82.2.399.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M. Chang, and Sally Grantham-McGregor. 2014. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344 (6187): 998–1001. https://doi.org/10.1126/science.1251178.

Grantham-McGregor, Sally, Christine Powell, Susan Walker, and John Himes. 1991. "Nutritional Supplementation, Psychosocial Stimulation, and Mental Development of Stunted Children: The Jamaican Study." *Lancet* 338: 1–5. https://doi.org/10.1016/0140-6736(91)90001-6.

Grantham-McGregor, Sally, Akanksha Adya, Orazio Attanasio, Britta Augsburg, Jere Behrman, Bet Caeyers, Monimalika Day, et al. 2020. "Group Sessions or Home Visits for Early Childhood Development in India: A Cluster RCT." *Pediatrics* 146 (6): 2020002725. http://publications.aap.org/pediatrics/article-

pdf/146/6/e2020002725/1241458/peds_2020002725.pdf.

Grantham-McGregor, Sally M., Lia C. H. Fernald, Rose M. C. Kagawa, and Susan Walker. 2014. "Effects of Integrated Child Development and Nutrition Interventions on Child Development and Nutritional Status." *Annals of the New York Academy of Sciences* 1308 (1): 11–32. https://doi.org/10.1111/nyas.12284.

Heckman, James, Bei Liu, Mai Lu, and Jin Zhou. 2020. "Treatment Effects and the Measurement of Skills in a Prototypical Home Visiting Program." Working Paper 27356, National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w27356.

Heckman, James J., and Stefano Mosso. 2014. "The Economics of Human Development and Social Mobility." *Annual Review of Economics* 6 (1): 689–733. https://doi.org/10.1146/annurev-economics-080213-040753.

Huntington-Klein, Nick. 2020. "MLRtime: A Stata package for running Machine Learning commands in R". https://github.com/NickCH-K/MLRtime/.

Hurley, Kristen M, Aisha K Yousafzai, and Florencia Lopez-boo. 2016. "Early Child Development and Nutrition : A Review of the Benefits and Challenges of." *Advances in Nutrition* 7 (2): 357–363. https://doi.org/10.3945/an.115.010363.by.

Jeong, Joshua, Emily E. Franchett, Clariana V. Ramos de Oliveira, Karima Rehmani, and Aisha K. Yousafzai. 2021. "Parenting Interventions to Promote Early Child Development in the First Three Years of Life: A Global Systematic Review and Meta-Analysis." *PLoS Medicine* 18 (5): e1003602. https://doi.org/10.1371/journal.pmed.1003602.

Johnson, Rucker C., and C. Kirabo Jackson. 2019. "Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending." *American Economic Journal: Economic Policy* 11 (4): 310–349. https://doi.org/10.1257/pol.20180510.

Knudsen, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff. 2006. "Economic, Neurobiological, and Behavioral Perspectives on Building

America's Future Workforce." *Proceedings of the National Academy of Sciences of the United States of America* 103 (27): 10155–10162.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. *Applied Predictive Modeling*. New York: Springer.

Luo, Renfu, Dorien Emmers, Nele Warrinnier, Scott Rozelle, and Sean Sylvia. 2019. "Using Community Health Workers to Deliver a Scalable Integrated Parenting Program in Rural China: A Cluster-Randomized Controlled Trial." *Social Science and Medicine* 239: 112545. https://doi.org/10.1016/j.socscimed.2019.112545.

Luo, Renfu, Ai Yue, Huan Zhou, Yaojiang Shi, Linxiu Zhang, Reynaldo Martorell, Alexis Medina, Scott Rozelle, and Sean Sylvia. 2017. "The Effect of a Micronutrient Powder Home Fortification Program on Anemia and Cognitive Outcomes among Young Children in Rural China: A Cluster Randomized Trial." *BMC Public Health* 17 (1): 1–16. https://doi.org/10.1186/s12889-017-4755-0.

Pérez-Escamilla, Rafael, and Victoria Hall Moran. 2017. "The Role of Nutrition in Integrated Early Child Development in the 21st Century: Contribution from the Maternal and Child Nutrition Journal." *Maternal and Child Nutrition* 13 (1): 1–4. https://doi.org/10.1111/mcn.12387.

Rao, Nirmala, Jin Sun, Eva E. Chen, and Patrick Ip. 2017. "Effectiveness of Early Childhood Interventions in Promoting Cognitive Development in Developing Countries: A Systematic Review and Meta-Analysis." *Hong Kong Journal of Paediatrics* 22 (1): 14–25.

Romano, Joseph P., and Michael Wolf. 2016. "Efficient Computation of Adjusted P-Values for Resampling-Based Stepdown Multiple Testing." *Statistics and Probability Letters* 113: 38–40. https://doi.org/10.1016/j.spl.2016.02.012.

Rubio-Codina, Marta, M. Caridad Araujo, Orazio Attanasio, Pablo Muñoz, and Sally Grantham-McGregor. 2016. "Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale

Studies." *PLoS ONE* 11 (8). https://doi.org/10.1371/journal.pone.0160962.

Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." *The Quarterly Journal of Economics* 70 (1): 65–94. https://doi.org/10.2307/1884513.

Sylvia, Sean, Nele Warrinnier, Renfu Luo, Ai Yue, Orazio Attanasio, Alexis Medina, and Scott Rozelle. 2021. "From Quantity to Quality: Delivering a Home-Based Parenting Intervention through China's Family Planning Cadres." *The Economic Journal* 131 (635): 1365–1400.

Tibshirani, Julie, Susan Athey, Rina Friedberg, Vitor Hadad, David Hirshberg, Luke Miner, Erik Sverdrup, Stefan Wager, Marvin Wright, and Maintainer Julie Tibshirani. 2022. "grf: Generalized Random Forests."

Walker, Susan P., Susan M. Chang, Christine A. Powell, and Sally M. Grantham-McGregor. 2004. "Psychosocial Intervention Improves the Development of Term Low-Birth-Weight Infants." *The Journal of Nutrition* 134 (6): 1417–1423. https://doi.org/10.1093/jn/134.6.1417.

Walker, Susan P., Susan M. Chang, Amika S. Wright, Rodrigo Pinto, James J. Heckman, and Sally M. Grantham-McGregor. 2021. "Cognitive, Psychosocial, and Behaviour Gains at Age 31 Years from the Jamaica Early Childhood Stimulation Trial." *Journal of Child Psychology and Psychiatry* 63 (3): 626–635. https://doi.org/10.1111/JCPP.13499.

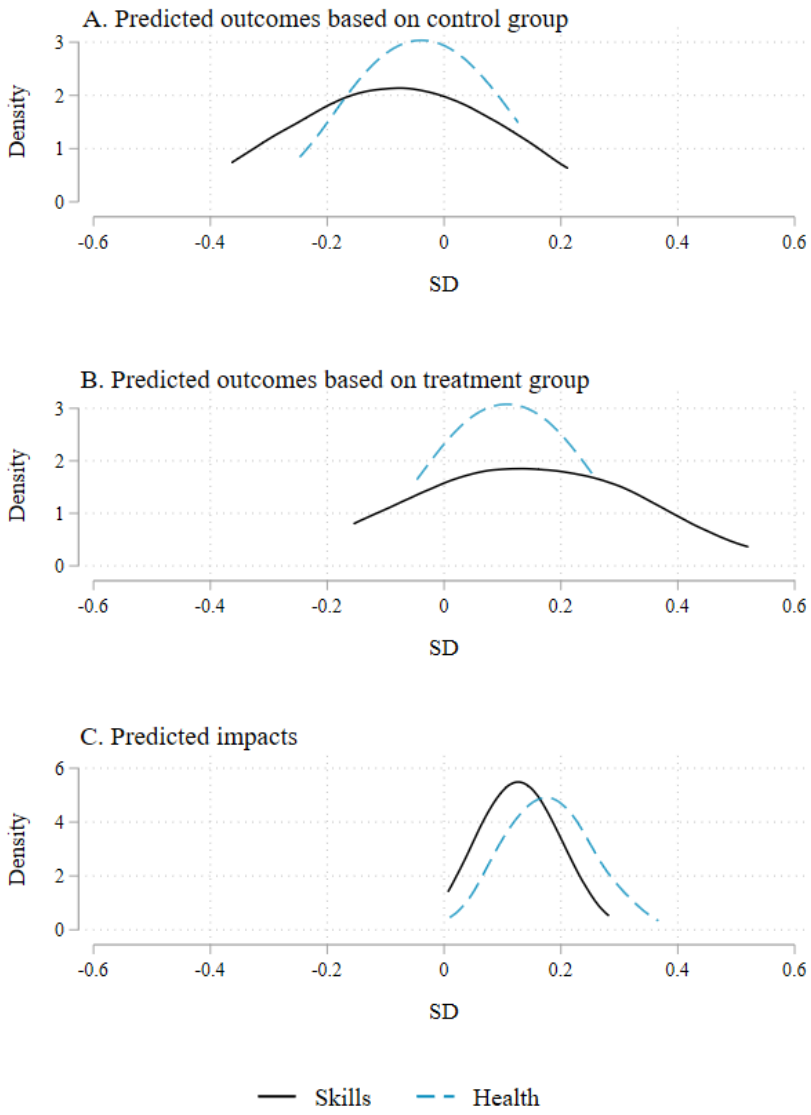
Wang, Lei, Wilson Liang, Siqi Zhang, Laura Jonsson, Mengjie Li, Cordelia Yu, Yonglei Sun, et al. 2019. "Are Infant/Toddler Developmental Delays a Problem across Rural China?" *Journal of Comparative Economics* 47 (2): 458–469. https://doi.org/10.1016/j.jce.2019.02.003.

Yousafzai, Aisha K., and Frances Aboud. 2014. "Review of Implementation Processes for Integrated Nutrition and Psychosocial Stimulation Interventions." *Annals of the New York Academy of Sciences* 1308 (1): 33–45. https://doi.org/10.1111/nyas.12313.

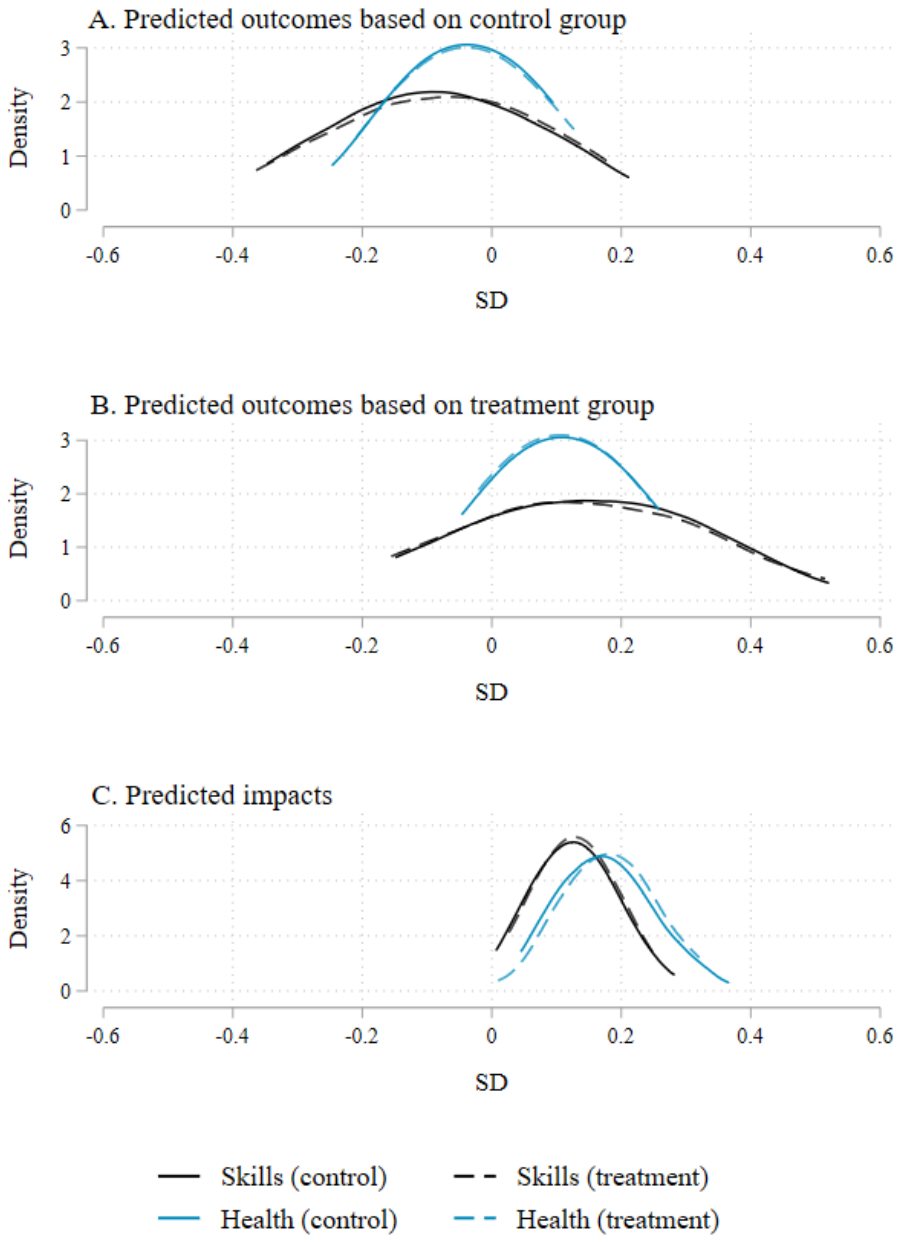
Appendix

Appendix Part A: Additional Figures and Tables

A.1: Additional Figures

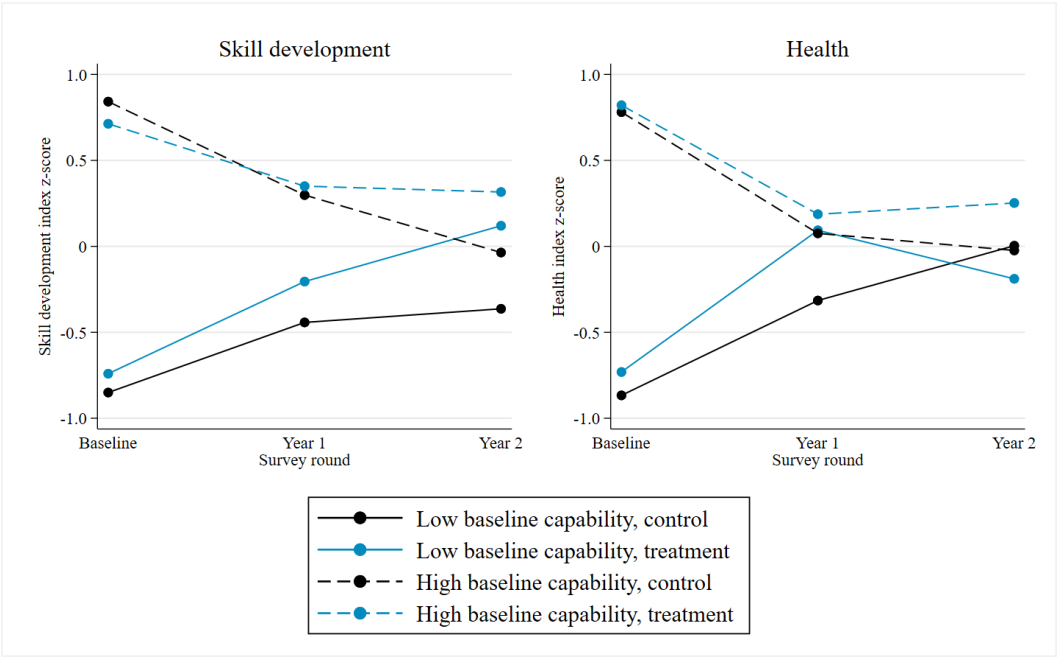


APPENDIX FIGURE A.1. KERNEL DENSITY PLOTS OF PREDICTED FIRST-YEAR OUTCOMES AND IMPACTS



APPENDIX FIGURE A.2. KERNEL DENSITY PLOTS OF PREDICTED FIRST-YEAR OUTCOMES AND IMPACTS BY TREATMENT ASSIGNMENT

Notes: T-test results show that the equality of means between control and treatment cannot be rejected at the 10% level of significance (with *p*-values ranging from .15 to .97). Kolmogorov-Smirnov tests show that the equality of the predicted outcome and impact distributions between treatment and control group cannot be rejected at the 10% level of significance (with *p*-values ranging from .22 to .92).



APPENDIX FIGURE A.3. CAPABILITY FORMATION OVER TIME

A.2: Additional Tables

APPENDIX TABLE A.1—TRAINER CHARACTERISTICS (*N*=36)

Variable	Mean	SD
Male	0.361	0.487
Married	0.944	0.232
Has children	0.972	0.167
Senior high school degree	0.417	0.500
Tertiary education degree	< 0.001	< 0.001
Village cadre	0.833	0.378
Government official	0.222	0.422
Farmer	0.611	0.494
Average travel time to households (in minutes)	19.083	11.382
Average travel distance to households (in km)	2.423	2.824

Source: Author calculations.

APPENDIX TABLE A.2—ITT EFFECTS ON ATTRITION

	Attrition Year 1	Attrition Year 2
	[1]	[2]
Treatment	0.016	-0.0002
	[0.045]	[0.050]
R^2	0.025	0.065
Observations	449	449
Control mean	0.119	0.247

Notes: Effects on attrition are estimated intervention effects controlling for province fixed effects. Standard errors [in square brackets] are adjusted for clustering at the village level and stratification at the township level.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.3—BASELINE BALANCE TABLE FOR STAYERS

	Full sample (1)	Control (2)	Treatment (3)	(2) vs. (3), *p*-value
A) Child characteristics				
(1) Age (in months)	12.845	12.972	12.711	.746
	(0.193)	*(0.309)*	*(0.226)*	
(2) Male	0.476	0.415	0.540	.027
	(0.027)	*(0.032)*	*(0.043)*	
(3) Born prematurely	0.045	0.035	0.055	.479
	(0.013)	*(0.015)*	*(0.022)*	
(4) First born	0.401	0.404	0.399	.793
	(0.027)	*(0.036)*	*(0.040)*	
(5) Cognition delay	0.485	0.450	0.521	.440
	(0.039)	*(0.058)*	*(0.047)*	
(6) Language delay	0.617	0.632	0.601	.512
	(0.039)	*(0.041)*	*(0.068)*	
(7) Motor delay	0.344	0.316	0.374	.629
	(0.035)	*(0.054)*	*(0.042)*	
(8) Social-emotional delay	0.518	0.509	0.528	.961
	(0.026)	*(0.036)*	*(0.038)*	
(9) Anemia (Hb < 110 g/L)	0.568	0.563	0.573	.568
	(0.035)	*(0.042)*	*(0.057)*	
(10) No. of times ill last month	0.844	0.906	0.779	.304
	(0.052)	*(0.064)*	*(0.079)*	
(11) Stunted (HAZ score < -2)	0.052	0.048	0.057	.939
	(0.013)	*(0.017)*	*(0.019)*	
(12) Wasted (WHZ score < -2)	0.018	0.018	0.019	.834
	(0.007)	*(0.010)*	*(0.010)*	
B) Household characteristics				
(1) Mom's age (in years)	28.324	28.240	28.414	.630
	(0.363)	*(0.411)*	*(0.620)*	
(2) Mother at home	0.892	0.889	0.896	.953
	(0.017)	*(0.019)*	*(0.029)*	
(3) Mother's education > 9 years	0.251	0.275	0.227	.433
	(0.025)	*(0.036)*	*(0.031)*	
(4) Household received social security support package	0.132	0.123	0.141	.633
	(0.017)	*(0.025)*	*(0.023)*	
Observations	334	171	163	

Notes: P-values account for clustering at the village level. In line with international standards, anemia is defined as an altitude-adjusted hemoglobin (*Hb*) concentration below 110 gram per liter (*g/L*); and stunting and wasting are defined as HAZ and WHZ scores below -2, respectively. Cognition, language, motor, and social-emotional delays are identified by a Bayley-III score that is more than 1 SD below the mean score of a healthy population. A joint significant test across all baseline characteristics in this table cannot reject the null hypothesis of balancedness across control and treatment at the 10% level.

Source: Author calculations.

APPENDIX TABLE A.4—ITT EFFECTS ON SUMMARY INDICES (WITH ADDITIONAL CONTROLS)

Year 1

	Primary outcomes			Secondary outcomes					
	Skill development index	Health index	Physical growth index	Parenting knowledge index	Psychosocial investment index	Home safety index	Dietary diversity index	Iron rich foods index	Micronutrient supplementation index
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Treatment	0.212***	0.243**	0.063	0.260***	0.384***	0.026	0.129	0.247**	0.101
	[0.079]	[0.102]	[0.079]	[0.093]	[0.087]	[0.142]	[0.100]	[0.115]	[0.100]
Observations	368	351	364	371	373	352	372	367	367
Control mean	-0.066	-0.123	-0.019	-0.127	-0.147	-0.019	-0.028	-0.093	-0.045
Additional controls	YES	YES	YES	YES	YES	YES	YES	YES	YES

Year 2

	Primary outcomes			Secondary outcomes					
	Skill development index	Health index	Physical growth index	Parenting knowledge index	Psychosocial investment index	Home safety index	Dietary diversity index	Iron rich foods index	Micronutrient supplementation index
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Treatment	0.321***	0.024	-0.072	0.413***	0.584***	0.346***	0.142	0.376***	0.498***
	[0.098]	[0.098]	[0.110]	[0.131]	[0.109]	[0.104]	[0.128]	[0.132]	[0.110]
Observations	308	274	294	315	313	286	314	308	309
Control mean	-0.190	0.004	0.027	-0.197	-0.246	-0.144	-0.055	-0.198	-0.253
Additional controls	YES	YES	YES	YES	YES	YES	YES	YES	YES

Notes: The (primary and secondary) outcome variables are expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. Effects on primary outcomes are estimated intervention effects controlling for baseline scores, province fixed effects, and additional controls. In column [1] we additionally controlled for tester fixed effects. The optimal set of additional controls (that minimizes the risk of omitted variable bias and overfitting) was selected using the "post-double-selection" (PDS) methodology of Belloni et al. (2016) implemented with Stata's pdslasso package. The pdslasso package estimates two lasso regressions to assess relationships of (1) the outcome variable of interest and (2) treatment assignment with the full set of controls. The final set of controls is limited to the union of controls that significantly affect the outcome of interest as well as treatment assignment. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.5—ITT EFFECTS ON SKILL DEVELOPMENT (IN *Z*-SCORES)

Year 1

	Skill development index *z*-score	Cognition	Receptive language	Expressive language	Gross motor	Fine motor	Social-emotional
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Treatment	0.181**	0.241*	-0.005	0.148	0.144	0.130	0.107
	[0.079]	[0.098]	[0.079]	[0.085]	[0.088]	[0.075]	[0.088]
	(0.027)	(0.019)	(0.953)	(0.090)	(0.110)	(0.093)	(0.234)
		{0.079}	{0.941}	{0.310}	{0.310}	{0.310}	{0.327}
R^2	0.333	0.230	0.228	0.258	0.311	0.25	0.199
Observations	385	386	386	386	385	386	386
Control mean	-0.066	-0.108	-0.023	-0.038	-0.093	-0.055	-0.02

Year 2

	Skill development index *z*-score	Cognition	Receptive language	Expressive language	Gross motor	Fine motor	Social-emotional
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Treatment	0.333***	0.283*	0.115	0.082	0.219	0.367**	0.150
	[0.101]	[0.117]	[0.085]	[0.088]	[0.119]	[0.127]	[0.121]
	(0.002)	(0.020)	(0.183)	(0.355)	(0.073)	(0.006)	(0.221)
		{0.089}	{0.317}	{0.317}	{0.188}	{0.020}	{0.317}
R^2	0.260	0.215	0.249	0.314	0.305	0.234	0.221
Observations	324	330	330	330	330	330	325
Control mean	-0.190	-0.086	-0.114	-0.077	-0.183	-0.190	-0.069

Notes: The skill development index is constructed as a weighted mean of the standardized values of the outcome variables in columns [2]-[7]. The outcome measures are expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. Effects on primary outcomes are estimated intervention effects controlling for baseline scores and province and tester fixed effects. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level. *P*-values in (parentheses) are uncorrected *p*-values. *P*-values in {curly brackets} are adjusted for multiple hypothesis testing using the Romano-Wolf stepdown procedure. Significance is determined based on *p*-values corrected for the family-wise error rate using the Romano-Wolf stepdown procedure**.**

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.6—ITT EFFECTS ON SKILL DEVELOPMENT (IN RAW SCORES)

Year 1

	Skill development index score	Cognition	Receptive language	Expressive language	Gross motor	Fine motor	Social-emotional
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Treatment	4.479**	1.285**	-0.107	0.666	0.459	0.228	1.75
	[1.949]	[0.511]	[0.290]	[0.387]	[0.304]	[0.256]	[1.264]
	(0.027)	(0.016)	(0.715)	(0.093)	(0.138)	(0.378)	(0.174)
		{0.030}	{0.693}	{0.287}	{0.346}	{0.545}	{0.347}
R^2	0.666	0.488	0.499	0.538	0.563	0.545	0.305
Observations	385	386	386	386	385	386	386
Control mean	314.337	60.812	26.360	28.543	57.832	40.066	100.848

Year 2

	Skill development index score	Cognition	Receptive language	Expressive language	Gross motor	Fine motor	Social-emotional
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Treatment	9.196**	1.507**	0.498	0.362	0.701	1.649**	3.307
	[3.410]	[0.623]	[0.549]	[0.517]	[0.395]	[0.644]	[2.551]
	(0.010)	(0.020)	(0.370)	(0.487)	(0.084)	(0.014)	(0.202)
		{0.050}	{0.426}	{0.426}	{0.178}	{0.050}	{0.366}
R^2	0.37	0.354	0.359	0.358	0.432	0.415	0.194
Observations	325	331	331	331	331	331	326
Control mean	395.458	74.521	37.077	39.296	64.769	50.456	129.491

Notes: The skill development index score is constructed as the sum of the raw Bayley-III subscale scores in columns [2]-[7]. The outcome measures are raw aggregated item scores. Effects on primary outcomes are estimated intervention effects controlling for baseline scores and province and tester fixed effects. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level. *P*-values in (parentheses) are uncorrected *p*-values. *P*-values in {curly brackets} are adjusted for multiple hypothesis testing using the Romano-Wolf stepdown procedure. Significance is determined based on *p*-values corrected for the family-wise error rate using the Romano-Wolf stepdown procedure.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.7—ITT EFFECTS ON HEALTH

Year 1					
	Health index *z*-score	Hemoglobin level (g/L)	Diarrhea	Respiratory Tract Infection	No. of times ill
	[1]	[2]	[3]	[4]	[5]
Treatment	0.249**	-0.218	-0.081***	-0.080	-0.081
	[0.107]	[1.850]	[0.021]	[0.048]	[0.071]
	(0.025)	(0.907)	(0.000)	(0.103)	(0.277)
		{0.891}	{0.010}	{0.208}	{0.356}
R^2	0.049	0.011	0.024	0.018	0.020
Observations	367	367	390	390	390
Control mean	-0.123	116.484	0.120	0.485	0.725
Year 2					
	Health index *z*-score	Hemoglobin level (g/L)	Diarrhea	Respiratory Tract Infection	No. of times ill
	[1]	[2]	[3]	[4]	[5]
Treatment	-0.063	1.509	-0.027	0.040	-0.050
	[0.097]	[1.598]	[0.029]	[0.049]	[0.068]
	(0.784)	(0.351)	(0.355)	(0.420)	(0.466)
		{0.614}	{0.614}	{0.614}	{0.614}
R^2	0.010	0.014	0.013	0.012	0.004
Observations	329	292	332	333	330
Control mean	0.028	116.104	0.047	0.541	0.692

Notes: The health index is constructed as a weighted mean of the standardized values of the outcome variables in columns [2]-[5]. This index score is expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. Hemoglobin levels are expressed in g/L and adjusted for altitude. The outcome variables in columns [3]-[4] are dummy indicators that take on the value of 1 if the child had this type of illness over the past two weeks, and 0 otherwise. Effects on primary outcomes are estimated intervention effects controlling for baseline scores and province fixed effects. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level. *P*-values in (brackets) are uncorrected *p*-values. *P*-values in {curly brackets} are adjusted for multiple hypothesis testing using the Romano-Wolf stepdown procedure. Significance is determined based on *p*-values corrected for the family-wise error rate using the Romano-Wolf stepdown procedure.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.8—ITT EFFECTS ON PHYSICAL GROWTH

Year 1					
	Physical growth index z-score	HAZ	Healthy HAZ	Healthy WHZ	Healthy WAZ
	[1]	[2]	[3]	[4]	[5]
Treatment	0.066	0.016	0.014	0.022	-0.016
	[0.078]	[0.089]	[0.018]	[0.025]	[0.015]
	(0.406)	(0.861)	(0.445)	(0.387)	(0.305)
		{0.772}	{0.594}	{0.594}	{0.554}
R^2	0.237	0.515	0.085	0.092	0.331
Observations	380	380	380	380	381
Control mean	-0.019	-0.074	0.945	0.920	0.940
Year 2					
	Physical growth index z-score	HAZ	Healthy HAZ	Healthy WHZ	Healthy WAZ
	[1]	[2]	[3]	[4]	[5]
Treatment	-0.119	0.113	0.014	-0.040	-0.04
	[0.120]	[0.157]	[0.042]	[0.036]	[0.033]
	(0.330)	(0.479)	(0.742)	(0.279)	(0.235)
		{0.594}	{0.752}	{0.455}	{0.386}
R^2	0.162	0.290	0.115	0.089	0.112
Observations	309	310	310	309	310
Control mean	0.027	-0.580	0.884	0.909	0.933

Notes: The physical growth index is constructed as a weighted mean of the standardized values of the outcome variables in columns [2]-[5]. This index score is expressed in z-scores. Z-scores are obtained by non-parametrically standardizing within age (in month) groups. HAZ, WHZ, and WAZ scores are calculated in line with the WHO standards. The outcome variables in columns [3]-[5] are dummy indicators that take on the value of 1 if the child has a healthy HAZ, WHZ, or WAZ score according to the WHO standards, and 0 otherwise. Effects on primary outcomes are estimated intervention effects controlling for baseline scores and province fixed effects. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level. *P*-values in (brackets) are uncorrected *p*-values. *P*-values in {curly brackets} are adjusted for multiple hypothesis testing using the Romano-Wolf stepdown procedure. Significance is determined based on *p*-values corrected for the family-wise error rate using the Romano-Wolf stepdown procedure.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.9—ITT EFFECTS ON PARENTING KNOWLEDGE AND BELIEFS

Year 1					
	Knowledge index z-score	Knows importance of reading	Knows importance of play	Able to read	Able to play
	[1]	[2]	[3]	[4]	[5]
Treatment	0.255***	0.070	0.063	0.117***	0.050
	[0.090]	[0.047]	[0.038]	[0.049]	[0.048]
	(0.007)	(0.145)	(0.102)	(0.020)	(0.310)
		{0.119}	{0.119}	{0.010}	{0.238}
R^2	0.048	0.030	0.030	0.015	0.048
Observations	388	389	389	390	390
Control mean	-0.127	0.51	0.735	0.435	0.635
Year 2					
	Knowledge index z-score	Knows importance of reading	Knows importance of play	Able to read	Able to play
	[1]	[2]	[3]	[4]	[5]
Treatment	0.388***	0.192***	0.094*	0.119**	0.137**
	[0.133]	[0.058]	[0.049]	[0.056]	[0.058]
	(0.006)	(0.002)	(0.065)	(0.038)	(0.023)
		{0.010}	{0.069}	{0.050}	{0.050}
R^2	0.042	0.047	0.026	0.015	0.026
Observations	331	332	332	333	333
Control mean	-0.197	0.535	0.7	0.524	0.6

Notes: The knowledge index is constructed as a weighted mean of the standardized values of the outcome variables in columns [2]-[5]. This index score is expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. The outcome variables in columns [2]-[5] are dummy indicators that take on the value of 1 if the caregiver holds this believe or has this ability, and 0 otherwise. Effects on primary outcomes are estimated intervention effects controlling for baseline scores and province fixed effects. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level. *P*-values in (brackets) are uncorrected *p*-values. *P*-values in {curly brackets} are adjusted for multiple hypothesis testing using the Romano-Wolf stepdown procedure. Significance is determined based on *p*-values corrected for the family-wise error rate using the Romano-Wolf stepdown procedure.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.10—ITT EFFECTS ON PSYCHOSOCIAL INVESTMENT

Year 1							
	Psychosocial investment index *z*-score	Play with toys	Tell stories	Read books	Sing songs	No. of books	Play area
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Treatment	0.352***	0.138**	0.094*	0.080	0.051	1.219	0.113**
	[0.085]	[0.060]	[0.049]	[0.051]	[0.037]	[0.736]	[0.040]
	(<0.001)	(0.026)	(0.064)	(0.122)	(0.181)	(0.105)	(0.007)
		{0.050}	{0.090}	{0.158}	{0.158}	{0.158}	{0.040}
R^2	0.235	0.062	0.064	0.028	0.085	0.104	0.109
Observations	390	390	390	390	390	390	390
Control mean	-0.147	0.435	0.16	0.09	0.31	3.36	0.295
Year 2							
	Psychosocial investment index *z*-score	Play with toys	Tell stories	Read books	Sing songs	No. of books	Play area
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Treatment	0.568***	0.198***	0.264***	0.240***	0.118**	3.687***	0.143***
	[0.103]	[0.043]	[0.056]	[0.041]	[0.053]	[1.097]	[0.048]
	(<0.001)	(<0.001)	(<0.001)	(0.002)	(0.032)	(0.002)	(0.005)
		{0.010}	{0.010}	{0.010}	{0.030}	{0.010}	{0.010}
R^2	0.201	0.07	0.093	0.107	0.04	0.075	0.105
Observations	329	333	333	333	331	332	332
Control mean	-0.246	0.341	0.182	0.118	0.335	6.665	0.376

Notes: The psychosocial investment index is constructed as a weighted mean of the standardized values of the outcome variables in columns [2]-[7]. This index score is expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. The outcome variables in columns [2]-[5] are dummy indicators that take on the value of 1 if the child was engaged in this activity during the previous day, and 0 otherwise. The outcome variable in column [7] is a dummy indicator that takes on the value of 1 if the child has its own play area, and 0 otherwise. Effects on primary outcomes are estimated intervention effects controlling for baseline scores and province fixed effects. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level. *P*-values in (brackets) are uncorrected *p*-values. *P*-values in {curly brackets} are adjusted for multiple hypothesis testing using the Romano-Wolf stepdown procedure. Significance is determined based on *p*-values corrected for the family-wise error rate using the Romano-Wolf stepdown procedure.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.11—ITT EFFECTS ON HOME SAFETY

Year 1					
	Home safety index *z*-score	Own play area	Smoking in home	Safety equipment installed	Dangerous objects removed
	[1]	[2]	[3]	[4]	[5]
Treatment	0.035	0.113**	0.046	-0.066	-0.03
	[0.142]	[0.040]	[0.041]	[0.053]	[0.033]
	(0.740)	(0.007)	(0.267)	(0.222)	(0.368)
		{0.020}	{0.376}	{0.376}	{0.376}
R^2	0.063	0.109	0.216	0.019	0.020
Observations	368	390	390	369	389
Control mean	-0.019	0.295	0.32	0.396	0.955
Year 2					
	Home safety index *z*-score	Own play area	Smoking in home	Safety equipment installed	Dangerous objects removed
	[1]	[2]	[3]	[4]	[5]
Treatment	0.293**	0.143**	<0.001	0.172**	0.009
	[0.112]	[0.048]	[0.055]	[0.065]	[0.019]
	(0.013)	(0.005)	(1.000)	(0.011)	(0.639)
		{0.020}	{1.000}	{0.030}	{0.772}
R^2	0.042	0.105	0.117	0.035	0.005
Observations	302	332	331	306	332
Control mean	-0.144	0.376	0.399	0.273	0.959

Notes: The home safety index is constructed as a weighted mean of the standardized values of the outcome variables in columns [2]-[5]. This index score is expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. The outcome variables in columns [2]-[5] are dummy indicators that take on the value of 1 if this protective or harmful environmental factor was present in the child's living environment, and 0 otherwise. Effects on primary outcomes are estimated intervention effects controlling for baseline scores and province fixed effects. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level. *P*-values in (brackets) are uncorrected *p*-values. *P*-values in {curly brackets} are adjusted for multiple hypothesis testing using the Romano-Wolf stepdown procedure. Significance is determined based on *p*-values corrected for the family-wise error rate using the Romano-Wolf stepdown procedure.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.12—ITT EFFECTS ON DIETARY DIVERSITY

Year 1										
	Minimum dietary diversity [1]	Iron rich foods index *z*-score [2]	Dietary diversity index *z*-score [3]	Grains, roots and tubers [4]	Legumes and nuts [5]	Dairy products [6]	Flesh foods [7]	Eggs [8]	Vitamin-A rich fruits or vegetables [9]	Other fruits and vegetables [10]
Treatment	0.033	0.231**	0.125	-0.005	0.017	0.008	0.026	-0.054	0.028	0.072
	[0.032]	[0.111]	[0.099]	[0.011]	[0.054]	[0.050]	[0.043]	[0.041]	[0.040]	[0.046]
	(0.304)	(0.044)	(0.216)	(0.615)	(0.762)	(0.866)	(0.551)	(0.201)	(0.493)	(0.126)
				{0.941}	{0.941}	{0.941}	{0.941}	{0.446}	{0.911}	{0.347}
R^2	0.027	0.066	0.089	0.002	0.003	0.125	0.025	0.06	0.018	0.048
Observations	389	384	389	389	389	389	389	389	389	389
Control mean	0.85	-0.093	-0.028	0.99	0.595	0.515	0.695	0.635	0.735	0.82
Year 2										
	Minimum dietary diversity [1]	Iron rich foods index *z*-score [2]	Dietary diversity index *z*-score [3]	Grains, roots and tubers [4]	Legumes and nuts [5]	Dairy products [6]	Flesh foods [7]	Eggs [8]	Vitamin-A rich fruits or vegetables [9]	Other fruits and vegetables [10]
Treatment	0.036	0.387***	0.141	0.009	0.004	-0.017	0.081	0.06	0.043	0.004
	[0.042]	[0.133]	[0.126]	[0.016]	[0.057]	[0.052]	[0.050]	[0.061]	[0.030]	[0.034]
	(0.403)	(0.006)	(0.270)	(0.578)	(0.940)	(0.746)	(0.110)	(0.328)	(0.165)	(0.912)
				{0.921}	{0.980}	{0.950}	{0.366}	{0.733}	{0.505}	{0.980}
R^2	0.004	0.044	0.013	0.009	0.007	0.067	0.048	0.068	0.031	0.002
Observations	330	323	330	332	332	331	332	332	332	331
Control mean	0.899	-0.198	-0.055	0.965	0.559	0.657	0.759	0.512	0.888	0.899

Notes: In line with the WHO criterion, the minimum dietary diversity indicator takes the value of 1 if the child received foods from 4 or more of the food groups in columns [4]-[10], and 0 otherwise. The iron rich foods index is the weighted average of the measures in columns [5], [6], [9], and [10] of this table and column [2] of Appendix Table A.13. The dietary diversity index is constructed as a weighted mean of the standardized values of the outcome variables in columns [4]-[10]. The iron rich foods index and dietary diversity index scores are expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. The outcome variables in columns [4]-[10] are dummy indicators that take the value of 1 if the child consumed foods from this food group, and 0 otherwise. Effects on primary outcomes are estimated intervention effects controlling for baseline scores and province fixed effects. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level. *P*-values in (brackets) are uncorrected *p*-values. *P*-values in {curly brackets} are adjusted for multiple hypothesis testing using the Romano-Wolf stepdown procedure. Significance is determined based on *p*-values corrected for the family-wise error rate using the Romano-Wolf stepdown procedure.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.13—ITT EFFECTS ON MICRONUTRIENT SUPPLEMENTATION

Year 1				
	Micronutrient supplementation index *z*-score	Iron	Zinc	Calcium
	[1]	[2]	[3]	[4]
Treatment	0.122	0.043	0.065	0.035
	[0.100]	[0.041]	[0.053]	[0.046]
	(0.229)	(0.309)	(0.226)	(0.455)
		{0.366}	{0.327}	{0.446}
R^2	0.077	0.01	0.048	0.108
Observations	384	384	385	389
Control mean	-0.045	0.19	0.32	0.58
Year 2				
	Micronutrient supplementation index *z*-score	Iron	Zinc	Calcium
	[1]	[2]	[3]	[4]
Treatment	0.505***	0.259***	0.203***	0.207***
	[0.109]	[0.053]	[0.058]	[0.044]
	(<0.001)	(<0.001)	(0.001)	(<0.001)
		{0.010}	{0.010}	{0.010}
R^2	0.079	0.068	0.044	0.137
Observations	324	324	328	332
Control mean	-0.253	0.406	0.464	0.629

Notes: The micronutrient supplementation index is constructed as a weighted mean of the standardized values of the outcome variables in columns [2]-[4]. This index score is expressed as a standardized *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. The outcome variables in columns [2]-[4] are dummy indicators that take on the value of 1 if the child consumed this supplement, and 0 otherwise. Effects on primary outcomes are estimated intervention effects controlling for baseline scores and province fixed effects. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level. *P*-values in (brackets) are uncorrected *p*-values. *P*-values in {curly brackets} are adjusted for multiple hypothesis testing using the Romano-Wolf stepdown procedure. Significance is determined based on *p*-values corrected for the family-wise error rate using the Romano-Wolf stepdown procedure.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.14—ITT EFFECTS ON NURSERY SCHOOL ENROLMENT

Panel A: Nursery school enrolment

	Full sample	Yunnan Province	Hebei Province
	[1]	[2]	[3]
Treatment	-0.078	-0.071	-0.050
	[0.055]	[0.090]	[0.070]
R^2	0.338	0.151	0.348
Observations	333	135	198
Control mean	0.56	0.30	0.71

Panel B: Nursery school enrolment age (in months)

	Full sample	Yunnan Province	Hebei Province
	[1]	[2]	[3]
Treatment	-0.545	0.453	-1.214*
	[0.434]	[0.462]	[0.704]
R^2	0.575	0.599	0.580
Observations	334	135	199
Control mean	35.60	34.55	36.16

Notes: Nursery school enrolment is a dummy indicator that takes the value of 1 if the child started attending nursery school by endline; and 0 otherwise. Effects on nursery school enrolment and nursery school enrolment age (in months) are estimated intervention effects controlling for baseline skill development scores, child sex, age, and province fixed effects. Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.15—VARIABLE IMPORTANCE

Baseline characteristics	As predictors of					
	Skill development index score		Impact on skill development	Health index score		Impact on health
	Control	Treatment		Control	Treatment	
	(1)	(2)	(3)	(4)	(5)	(6)
Skill development index *z*-score	45.0%	36.8%	12.2%	5.4%	7.3%	8.0%
Home safety index *z*-score	2.2%	4.7%	9.8%	5.8%	5.7%	7.2%
Psychosocial investment index *z*-score	6.3%	8.2%	7.6%	12.5%	7.7%	10.1%
Iron rich foods index *z*-score	5.9%	5.3%	8.5%	7.3%	11.9%	9.1%
Nutrition supplementation index *z*-score	4.3%	2.9%	6.5%	17.4%	8.0%	16.1%
Health index *z*-score	4.8%	3.4%	7.0%	17.7%	16.3%	8.6%
Physical growth index *z*-score	3.8%	2.9%	6.3%	2.7%	11.3%	7.8%
Parenting knowledge index *z*-score	3.6%	6.8%	11.0%	7.8%	12.2%	10.5%
Hemoglobin value (g/L)	3.8%	6.0%	13.3%	7.1%	6.7%	6.2%
Wealth index *z*-score	12.8%	13.8%	8.6%	11.3%	4.5%	10.7%
Child's sex (1=male; 0=female)	2.1%	0.4%	3.2%	1.1%	0.5%	0.8%
Child's age (in months)	1.7%	2.7%	5.3%	3.5%	5.4%	3.7%
Province	3.9%	6.1%	0.5%	0.5%	2.5%	1.2%

Notes: The variable importance is the frequency with which each observable baseline characteristic is used as a splitting variable in the (generalized) random forest algorithm.

Source: Author calculations.

APPENDIX TABLE A.16—K-FOLD CROSS-VALIDATION TO OPTIMIZE OUT-OF-SAMPLE PREDICTION PERFORMANCE

	Skill development index z-score				Health index z-score			
	Random forest	OLS	Lasso	Square-root lasso	Random forest	OLS	Lasso	Square-root lasso
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
K = 2								
Control	0.719	0.801	0.893	0.908	1.203	1.374	1.164	1.164
	0.110	0.132	0.140	0.142	0.199	0.210	0.184	0.184
Treatment	0.747	1.035	0.982	0.964	0.810	0.942	0.801	0.801
	0.113	0.167	0.146	0.142	0.137	0.149	0.140	0.140
K = 4								
Control	0.692	0.727	0.830	0.873	1.173	1.314	1.165	1.165
	0.147	0.176	0.184	0.192	0.281	0.280	0.268	0.268
Treatment	0.783	0.969	0.951	0.990	1.585	1.780	1.583	1.583
	0.236	0.266	0.270	0.289	0.350	0.395	0.363	0.363
K = 8								
Control	0.721	0.824	0.874	0.915	1.186	1.330	1.199	1.199
	0.221	0.255	0.269	0.279	0.398	0.420	0.394	0.394
Treatment	0.790	1.012	0.943	0.972	0.799	0.860	0.801	0.801
	0.248	0.302	0.283	0.292	0.259	0.283	0.270	0.270

Notes: We report the *K*-fold cross-validation mean squared prediction error, which is constructed as follows.

$$CV_K^{MSPE} = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{n_k}\sum_{i\in K_k}\left(y_i - x_i'\hat{\beta}_{-k}\right)^2\right)$$

Predictors with variable importance are listed in Appendix Table A.15.

Source: Author calculations.

APPENDIX TABLE A.17—HETEROGENEOUS ITT EFFECTS ON HEALTH AT END OF YEAR 2 BY BASELINE OR PREDICTED SKILL DEVELOPMENT AND/OR HEALTH AT END OF YEAR 1

	Dependent variable: Health index *z*-score at end of year 2								
	Subgroups: Low vs. high skill development			Subgroups: Low vs. high health			Subgroups: Low vs. high skills and health		
	At baseline	Year 1 prediction based on control group	Year 1 prediction based on treatment group	At baseline	Year 1 prediction based on control group	Year 1 prediction based on treatment group	At baseline	Year 1 prediction based on control group	Year 1 prediction based on treatment group
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Estimated ITT effect									
Low outcome group	0.087	-0.016	-0.069	-0.232	-0.051	-0.070	-0.339*	-0.012	-0.136
	0.12	0.148	0.14	0.151	0.177	0.123	0.178	0.196	0.157
High outcome group	-0.041	0.067	0.133	0.288*	0.110	0.132	0.045	0.166	0.167
	0.143	0.135	0.148	0.168	0.106	0.148	0.272	0.126	0.215
Health index *z*-score at baseline	0.143**	0.142**	0.143**	0.185*	0.145**	0.146**	0.224	0.186***	0.222***
	0.063	0.061	0.061	0.093	0.06	0.061	0.188	0.068	0.078
R^2	0.021	0.020	0.023	0.037	0.022	0.027	0.04	0.037	0.049
Observations	288	288	288	288	288	288	123	189	179
P-value of Wald test for homogeneity of estimated ITT effects	0.497	0.692	0.333	0.057	0.454	0.305	0.215	0.437	0.212

Notes: Health index scores are expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. The treatment effects on child health in columns [1]-[9] are regression coefficients on the treatment variable by baseline/predicted subgroup in a linear regression that includes controls for the baseline/predicted outcome groups, baseline health, and province fixed effects. Skill development and health outcomes at the end of the first intervention year are predicted using a random forest algorithm. Standard errors are computed with the cluster-correlated Huber-White estimator with clustering at the village level.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.18—HETEROGENEOUS ITT EFFECTS ON SKILL DEVELOPMENT AND HEALTH AT END OF YEAR 2 BY PREDICTED IMPACT ON SKILL DEVELOPMENT OR HEALTH AT END OF YEAR 1

	Dependent variable: Skill development index *z*-score at end of year 2		Dependent variable: Health index *z*-score at end of year 2	
	Subgroups: Low vs. high predicted year 1 impact on skills [1]	Subgroups: Low vs. high predicted year 1 impact on health [2]	Subgroups: Low vs. high predicted year 1 impact on skills [3]	Subgroups: Low vs. high predicted year 1 impact on health [4]
Estimated ITT effect				
Low predicted impact group	0.390***	0.482***	0.099	0.017
	0.133	0.174	0.139	0.131
High predicted impact group	0.294*	0.25	0.058	0.124
	0.150	0.154	0.149	0.136
Dependent variable at baseline	0.232***	0.209***	0.118**	0.121**
	0.051	0.058	0.059	0.058
R^2	0.258	0.265	0.025	0.024
Observations	300	295	276	276
P-value of Wald test for homogeneity of estimated ITT effects	0.594	0.311	0.85	0.574

Notes: Skill development and health index scores are expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. The treatment effects on child skill development or health in columns [1]-[4] are regression coefficients on the treatment variable by predicted subgroup in a linear regression that includes controls for the predicted impact groups, baseline skill development, and province fixed effects. Impacts on skill development and health outcomes at the end of the first intervention year are predicted using the generalized random forest algorithm of Athey, Tibshirani, and Wager (2019). Standard errors are computed with the cluster-correlated Huber-White estimator with clustering at the village level.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE A.19—HETEROGENEOUS ITT EFFECTS ON PARENTAL INVESTMENT IN CHILD PSYCHOSOCIAL INVESTMENT AT END OF YEAR 2 BY BASELINE OR PREDICTED SKILL DEVELOPMENT AND/OR HEALTH AT END OF YEAR 1

	Dependent variable: Psychosocial investment index *z*-score at end of year 2								
	Subgroups: Low vs. high skill development			Subgroups: Low vs. high health			Subgroups: Low vs. high skills and health		
	At baseline	Year 1 prediction based on control group	Year 1 prediction based on treatment group	At baseline	Year 1 prediction based on control group	Year 1 prediction based on treatment group	At baseline	Year 1 prediction based on control group	Year 1 prediction based on treatment group
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Estimated ITT effect									
Low outcome group	0.571***	0.693***	0.619***	0.762***	0.546***	0.690***	0.861***	0.693***	0.845***
	0.146	0.144	0.157	0.149	0.164	0.153	0.189	0.21	0.197
High outcome group	0.523***	0.437***	0.507***	0.346**	0.521***	0.462***	0.383**	0.517***	0.621***
	0.139	0.14	0.135	0.163	0.145	0.146	0.187	0.179	0.18
Psychosocial investment index *z*-score at baseline	0.313***	0.328***	0.327***	0.317***	0.304***	0.386***	0.336***	0.305***	0.382***
	0.064	0.063	0.062	0.062	0.062	0.079	0.094	0.089	0.082
R^2	0.238	0.243	0.24	0.244	0.242	0.252	0.300	0.200	0.240
Observations	327	327	327	317	327	327	140	211	204
P-value of Wald test for homogeneity of estimated ITT effects	0.787	0.161	0.554	0.067	0.911	0.269	0.108	0.500	0.385

Notes: Psychosocial investment index scores are expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. The treatment effects on child psychosocial stimulation in columns [1]-[9] are regression coefficients on the treatment variable by baseline/predicted subgroup in a linear regression that includes controls for the baseline/predicted outcome groups, baseline psychosocial investment, and province fixed effects. Skill development and health outcomes at the end of the first intervention year are predicted using a random forest algorithm. Standard errors are computed with the cluster-correlated Huber-White estimator with clustering at the village level.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Appendix Part B: Drivers and Impacts of Compliance

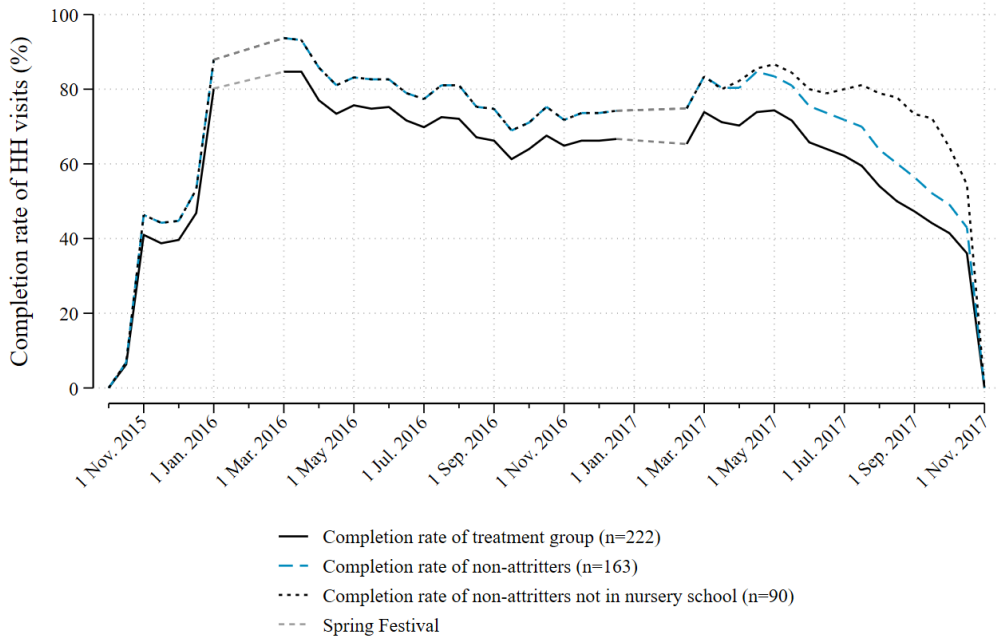
B.1: Estimation

Under full compliance, 24 bi-weekly home visits would have been delivered each year. As reported in Appendix Table B.1, on average, 29.3 home visits were delivered per child during the two-year intervention period. On average, 15.4 and 13.9 visits per child were delivered during the first and second year, respectively. Moreover, an average of 32.8 home visits were delivered per non-attritted child, of which 17.1 and 15.9 were delivered during the first and second year. As displayed in Appendix Figure B.1, no home visits were delivered during the Spring Festival. The completion rate of home visits for each bi-weekly time period was low (0%-60%) during the first three months of the intervention, but then centered around 80% for non-attritters. By the end of the intervention period average completion rates decreased, often because children reached three years of age and started attending nursery school (and so were not at home for the parental training sessions).

APPENDIX TABLE B.1—INTERVENTION ACTIVITIES

Total number of household visits delivered		
	2015-2016	3414
	2016-2017	3087
	2015-2017	6501
Average number of home visits delivered per child		
	2015-2016	15.4
	2016-2017	13.9
	2015-2017	29.3
Average number of home visits delivered per non-attritted child		
	2015-2016	17.1
	2016-2017	15.9
	2015-2017	32.8

Source: Author calculations.



APPENDIX FIGURE B.1. COMPLETION RATES OF HOME VISITS OVER TIME

In our analysis of the drivers and impacts of compliance, we pay attention to changes in compliance patterns and the marginal returns per home visit that can be explained by differences in child age and entry in nursery school. We start with the analysis of the drivers of compliance. We use a two-step procedure to measure the degree of correlation between the number of completed home visits and a number of different explanatory variables (including the age of the child). First, we correlate the number of completed home visits with a number of time-invariant determinants of compliance (i.e., determinants that explain compliance during the first and the second year). To be specific, we assess the relationship between the number of home visits and children's sex, baseline age, baseline cognition, and the travel distance between the home and the home of the parenting instructor. Second, we investigate how first-year experience drives compliance during the second intervention year. We study the role of caregiver's own experiences and the

experiences of peers in the village.[15] As shown in Appendix Table B.2, peers such as family members and friends are an important source of parenting information for the caregivers in our sample. To be precise, 60.1% and 38.3% of the caregivers in our sample obtained parenting information through family members and friends, respectively.[16]

APPENDIX TABLE B.2—PARENTING INFORMATION SOURCES OF CAREGIVERS

Family members	60.1%
Friends	38.3%
Internet	35.0%
TV	25.8%
Books	17.3%
Village doctor, representative of the NHC, representative of the Women's Federation, or another type of parenting expert	14.8%
None	11.4%
Observations	446

Notes: This is the percentage of caregivers who reported that they received parenting information via these sources.

Source: Author calculations.

Incomplete compliance may initially attenuate intervention impacts, but after the optimal number of home visits is reached, the returns per visit may decrease. In order to investigate the optimal number of home visits, we use a control function approach that allows for nonlinearity to estimate dose-response relationships. We first use OLS to estimate the first-stage relationship:

$$N_{ijt} = \alpha_0 + \alpha_1 T_j + \alpha_2 T_j * D_{ijt} + \alpha_3 D_{ijt} + \alpha_4 A_{ij2} + \alpha_5 Y_{ij0} + \pi_s + \eta_{ijt} \quad \text{(B.1)}$$

[15] We assess a caregiver's own first-year experience with the number of completed home visits and the average percentage gain in child skill development per completed home visit. We aggregate the number of first-year home visits and gains in skill development per visit across all caregivers in the village as measures of average peer experience. As explained in detail in, e.g., Angrist (2014), correlations between individual outcomes and group average outcomes need to be interpreted with care. Therefore, we additionally construct leave-out means (i.e., we exclude the individual's own outcome before computation of group average outcomes).

[16] In addition, 35.0%, 25.8%, and 17.3% of the caregivers obtained parenting information from the Internet, TV, and books, respectively (see Appendix Table B.2). 14.8% reported that they received information from village doctors, representatives of the NHC of the Women's Federation, or another type of parenting expert. 11.4% of the caregivers obtained no parenting information or advice from any type of external source.

We instrument the number of home visits delivered to child *i* in village *j* (N_{ijt}) during *t* years (t=1 or 2) with treatment assignment (T_j), the distance from the home to the home of the parenting instructor (D_{ijt}), the interaction between these two variables, an indicator for whether or not the child reaches the nursery school attendance age by endline (A_{ij2}), baseline skill development (Y_{ij0}), and province and tester fixed effects (π_s).[17] Standard errors are adjusted for clustering at the village level.

In the second stage, we use first-stage residual estimates ($\hat{\eta}_{ijt}$) as instruments to estimate the following OLS specifications, first assuming a linear relationship and then allowing for a nonlinear relationship by adding a squared term for the number of completed household visits:

$$Y_{ijt} = \alpha_0 + \alpha_1 N_{ijt} + \alpha_2 \hat{\eta}_{ijt} + \alpha_3 Y_{ij0} + \pi_s + \varepsilon_{ijt} \quad \text{(B.2)}$$

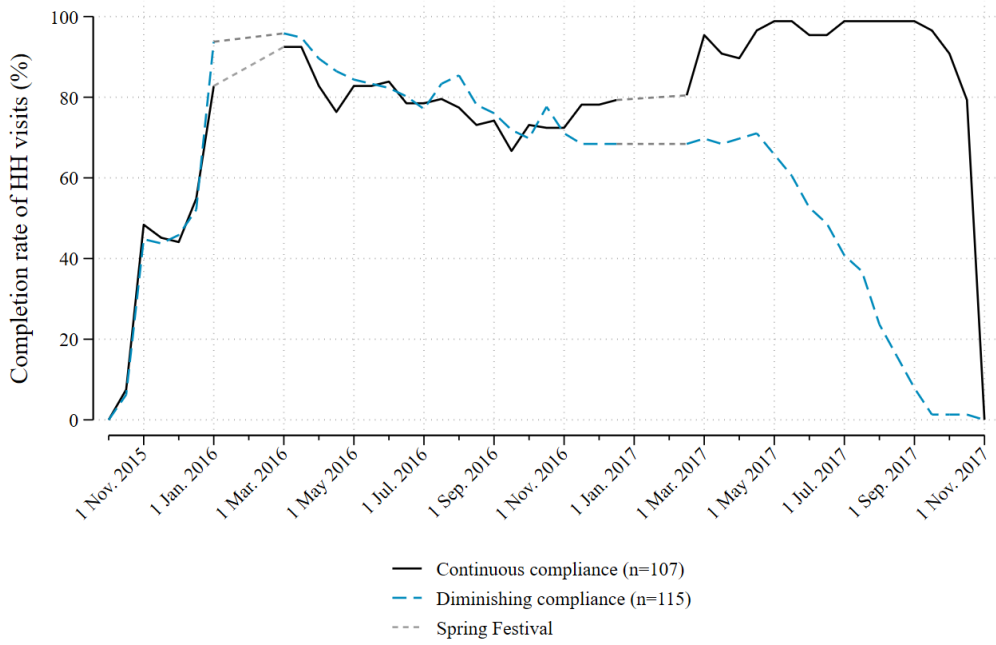
$$Y_{ijt} = \alpha_0 + \alpha_1 N_{ijt} + \alpha_2 N_{ijt}^2 + \alpha_3 \hat{\eta}_{ijt} + \alpha_4 \hat{\eta}_{ijt}^2 + \alpha_3 Y_{ij0} + \pi_s + \varepsilon_{ijt} \quad \text{(B.3)}$$

The notation in these specifications is in line with the previous specifications in this section. N_{ijt} and N_{ijt}^2 are the number and the squared number of home visits during *t* years, respectively. $\hat{\eta}_{ijt}$ and $\hat{\eta}_{ij2}^2$ are the estimated residual and the squared residual of the first-stage specification. We adjusted standard errors for clustering at the village level.

Finally, we investigate whether it is optimal to continue or end program delivery when children start attending nursery school. We assess two-year impact heterogeneity across two clusters of children: children who participate in the program during the full intervention period and children who leave the program over the course of the intervention period when they start attending nursery school. We apply Ward's minimum variance hierarchical clustering algorithm (Ward 1963)

[17] Note that the distance between the home and the home of the parenting instructor or reaching nursery school attendance age by endline are not significantly correlated with child skill development at baseline ($p > .10$).

to a list of bi-weekly indicator dummies for participation in home visits during the final six months of the intervention period (i.e., when part of the children reaches three years of age and start attending nursery school) to define two clusters of compliance: continuous compliance and diminishing compliance. The indicator dummies take the value of one if the child participates in the planned home visit during this bi-weekly period, and the value of 0 otherwise. We measure similarity with the simple matching binary similarity coefficient (Zubin 1938; Sokal and Michener 1958). Appendix Figure B.2 displays the completion rates of home visits over time for these two clusters of compliance.



APPENDIX FIGURE B.2. COMPLETION RATES OF HOME VISITS OVER TIME BY CLUSTERS OF COMPLIANCE

After defining these clusters, we conduct a traditional heterogeneity analysis in line with regression specification (5):

$$Y_{ij2} = \alpha_0 + \alpha_1 T_j C_{ij}^{CONT} + \alpha_2 T_j C_{ij}^{DIM} + \alpha_3 C_{ij}^{CONT} + \alpha_4 Y_{ij0} + \pi_s + \varepsilon_{ij} \qquad \text{(B.4)}$$

We regress capabilities of child *i* in village *j* at the end of the second year (Y_{ij2}) on dummy indicators ($T_j C_{ij}^{CONT}$ and $T_j C_{ij}^{DIM}$) indicating whether child *i* in treatment

group village *j* belonged to the cluster with continuous or diminishing compliance, respectively; the dummy indicator that takes the value of one if the child belongs to the group with continuous compliance (C_{ij}^{CONT}), and the value of zero otherwise; the child's capability score at baseline (Y_{ij0}), province fixed effects (π_s), and for treatment impacts on child skills also tester fixed effects.

B.2: Drivers of Compliance

Panel A of Appendix Table B.3 investigates the role of time-invariant drivers of compliance. Children's sex isn't significantly correlated with the number of completed home visits (columns 1 to 5, row 1). If the child is over 12 months of age at baseline, the child receives, on average, 2 to 3 visits less (columns 1 to 5, row 2). Children that are over 12 months of age at baseline, reach three years of age before the endline date of the trial. This implies that these children reach nursery school attendance age. If they start attending nursery school, they are likely to complete fewer home visits towards the end of the trial. They receive, on average, 2 to 3 more visits if their baseline cognition is below the median (columns 1 to 5, row 3). Children receive on average one visit less per kilometer distance between the home and the home of the parenting trainer (columns 1 to 5, row 4).

APPENDIX TABLE B.3—DRIVERS OF COMPLIANCE

	HH visits $(t_2 - t_0)$ [1]	HH visits $(t_2 - t_1)$ [2]	HH visits $(t_2 - t_1)$ [3]	HH visits $(t_2 - t_1)$ [4]	HH visits $(t_2 - t_1)$ [5]
Panel A: Time-invariant drivers of compliance					
Child is female	0.367	0.217	0.260	0.234	0.255
	(1.287)	(0.739)	(0.641)	(0.613)	(0.616)
Child's baseline age>12 months	-3.287*	-2.568*	-2.245*	-2.145*	-2.167*
	(1.850)	(1.254)	(1.249)	(1.235)	(1.239)
Low baseline cognition	2.309*	2.666***	2.790***	2.824***	2.820***
	(1.154)	(0.703)	(0.734)	(0.736)	(0.744)
Travel distance for trainer (km)	-1.179***	-1.074***	-1.064***	-1.050***	-1.039***
	(0.230)	(0.201)	(0.189)	(0.196)	(0.197)
Panel B: Time-variant drivers of compliance					
Own experience					
HH visits $(t_1 - t_0)$			0.285**	0.327**	0.316**
			(0.119)	(0.130)	(0.121)
Change in child skill development per HH visit $(t_1 - t_0)$			0.736***	0.584***	0.763***
			(0.210)	(0.205)	(0.219)
Experience of peers in village					
Mean of HH visits $(t_1 - t_0)$				-0.139	
				(0.204)	
Mean change in child skill development per HH visit $(t_1 - t_0)$				2.234*	
				(1.185)	
Leave-out mean of HH visits $(t_1 - t_0)$					-0.139
					(0.159)
Leave-out mean change in child skill development per HH visit $(t_1 - t_0)$					2.212*
					(1.223)
Province FE	YES	YES	YES	YES	YES
R^2	0.34	0.29	0.35	0.35	0.36
Observations	183	183	183	183	183

Notes: HH visits $(t_2 - t_0)$ = the number of home visits during the two years of the intervention (October 2015-September 2017). HH visits $(t_2 - t_1)$ = the number of home visits during the second year of the intervention (October 2016- September 2017). HH visits $(t_1 - t_0)$ = the number of home visits during the first year of the intervention (October 2015-September 2017). Child is female = dummy indicator that takes the value of 1 if the child is male, and 0 otherwise. Low baseline cognition = Bayley-III cognition raw score below the median. Change in skill development per HH visit = the skill growth rate divided by the number of home visits. Standard errors in (parentheses) are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Furthermore, Panel B of Appendix Table B.3 shows that a family's own experience and the experience of a family's peers during the first year drive the number of completed visits during the second year. Second-year home visits increase, on average, by 3 visits per 10 completed home visits in the first year (columns 3 to 5, row 5). In addition, the number of second-year home visits increases with 0.058-0.076 visits per 10% increase in skill development per home visit during year 1 (columns 3 to 5, row 6). Furthermore, if village peers perceive that standardized skill development scores increase on average by 10% per home visit during year 1, caregivers complete on average 0.22 home visits more during the second year (columns 4 and 5, rows 8 and 10).

Appendix Table B.4 provides evidence on the impact of compliance patterns on treatment impacts. The table presents the estimated ITT effects on summary indices of primary and secondary outcome measures by clusters of compliance. At the 10% significance level, we detect no differences in the treatment impacts between children with continuous or diminishing compliance towards the end of the intervention period, when children reach age 3 (i.e., the nursery school enrolment age). Given that treatment impacts on children with diminishing compliance are no smaller than treatment impacts on children with continued compliance, we argue that it is optimal to end program delivery when children enter nursery school.

APPENDIX TABLE B.4—HETEROGENEITY IN ITT EFFECTS ACROSS CLUSTERS OF COMPLIANCE

	Primary outcomes			Secondary outcomes					
	Skill development index [1]	Health index [2]	Physical growth index [3]	Parenting knowledge index [4]	Psychosocial investment index [5]	Home safety index [6]	Dietary diversity index [7]	Iron rich foods index [8]	Micronutrient supplementation index [9]
Estimated ITT effect									
Continuous compliance	0.291**	-0.048	-0.101	0.347***	0.638***	0.382***	0.138	0.427***	0.578***
	0.116	0.139	0.164	0.128	0.112	0.128	0.14	0.138	0.118
Diminishing compliance	0.386***	0.103	-0.137	0.435**	0.488***	0.178	0.144	0.340**	0.423***
	0.119	0.142	0.146	0.17	0.119	0.155	0.156	0.144	0.144
R^2	0.261	0.023	0.162	0.043	0.204	0.047	0.013	0.044	0.082
Observations	324	288	309	331	329	302	330	323	324
P-value of Wald test for homogeneity of estimated ITT effects	0.425	0.468	0.856	0.528	0.164	0.251	0.971	0.341	0.293

Notes: The (primary and secondary) outcome variables are expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. The treatment effects on skill development in columns [1]-[9] are regression coefficients on the treatment variable by cluster of compliance in a linear regression that includes controls for clusters of compliance, the dependent variable at baseline, and province and tester fixed effects. Clusters of compliance are produced by applying Ward's linkage hierarchical clustering algorithm (Ward 1963) using the simple matching binary similarity coefficient (Zubin 1938; Sokal and Michener 1958). Standard errors in [square brackets] are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

B.3: Dose-response Estimates

Appendix Table B.5 displays dose-response estimates for the sample.[18] In columns (1), we assume a linear relationship between the number of completed home visits and the key outcomes of interest (i.e., skill development and health index *z*-scores). Panels A and B report results of dose-response relationships for the first year and for both years of the intervention, respectively. The results in columns (1) of Panel A indicate that up to a point where individuals participate in 24 bi-weekly home visits over a one-year period, child skill development and health investment increased, on average, by 0.011 and 0.017 SD per home visit, respectively ($p < 0.05$). As shown in columns (2) of Panel A, we detect no evidence of concavity in the dose-response relationship with regard to skill development and health investment during the first year.

The two-year linear dose-response estimates in columns (1) of Panel B indicate that up to a point where individuals participate in 48 bi-weekly home visits over a two-year period, child skill development increased, on average, by 0.009 SD per home visit ($p < 0.01$). As shown in column (2), however, the analysis suggests that there is evidence of concavity in the dose-response relationship with regard to skill development. Our results indicate that skill development index *z*-scores increase, on average, by 0.076 SD per completed home visit, but scores decrease by -0.002 SD per squared number of visits ($p < 0.05$). We find no evidence of a significant two-year dose-response relationship between the number of visits and child health index *z*-scores.

[18] First-stage regression results are reported in Appendix Table B.6.

APPENDIX TABLE B.5—DOSE-RESPONSE RELATIONSHIPS

Panel A: First year

	Skill development index *z*-score		Health index *z*-score	
	[1]	[2]	[1]	[2]
HH visits	0.011**	-0.034	0.017**	0.147
	(0.005)	(0.108)	(0.007)	(0.108)
HH visits^2		0.002		-0.007
		(0.006)		(0.006)
Dependent variable at baseline	0.329***	0.332***	0.173***	0.174***
	(0.044)	(0.044)	(0.061)	(0.062)
Observations	385	385	367	367
R^2	0.333	0.334	0.050	0.053
Province FE	YES	YES	YES	YES
Tester FE	YES	YES	NO	NO

Panel B: Two years

	Skill development index *z*-score		Health index *z*-score	
	[1]	[2]	[1]	[2]
HH visits	0.009***	0.076***	0.002	0.024
	(0.003)	(0.027)	(0.003)	(0.027)
HH visits^2		-0.002**		-0.001
		(0.001)		(0.001)
Dependent variable at baseline	0.250***	0.229***	0.138**	0.144**
	(0.047)	(0.049)	(0.059)	(0.058)
Observations	324	324	288	288
R^2	0.257	0.269	0.026	0.032
Province FE	YES	YES	YES	YES
Tester FE	YES	YES	NO	NO

Notes: The outcome variables are expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. Column (1) gives the control function estimate of the treatment effect of one household visit, assuming a linear relationship between the number of household visits and the outcome variable of interest. Column (2) gives 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, the travel distance from the home of the parenting instructor to the home of the household, the interaction between treatment status and the distance, and an indicator variable for whether the child reaches the starting age for nursery school (i.e., three years of age) by endline. In all regressions we control for the dependent variable at baseline and province and tester fixed effects. Standard errors in (parentheses) are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

APPENDIX TABLE B.6—FIRST-STAGE REGRESSION RESULTS

	Year 1				Year 2			
	HH visits [1]	HH visits^2 [2]	HH visits [1]	HH visits^2 [2]	HH visits [1]	HH visits^2 [2]	HH visits [1]	HH visits^2 [2]
Treatment status	17.58***	342.0***	15.91***	307.7***	35.71***	1,355***	32.54***	1,226***
	(0.911)	(24.71)	(1.008)	(26.06)	(1.441)	(83.93)	(1.563)	(82.91)
Travel distance for trainer to visit home (km)	0.468***	14.81***	0.441***	13.17***	0.544***	33.18***	0.477**	26.68**
	(0.165)	(4.957)	(0.163)	(4.547)	(0.195)	(11.03)	(0.201)	(10.84)
Treatment * travel distance (km)	-0.657**	-21.27***	-0.529*	-17.30**	-1.881***	-111.5***	-1.739***	-105.8***
	(0.274)	(7.072)	(0.280)	(6.843)	(0.392)	(22.33)	(0.514)	(25.85)
Reaches nursery school starting age by endline	0.588*	11.13	0.486	8.487	-2.326***	-161.2***	-1.694***	-120.6***
	(0.321)	(7.762)	(0.338)	(6.999)	(0.858)	(53.80)	(0.563)	(35.71)
Skill development index *z*-score at baseline	-0.443**	-9.693*			-0.778**	-41.07*		
	(0.198)	(5.302)			(0.349)	(20.90)		
Health index *z*-score at baseline			0.340	6.333			0.758*	30.65*
			(0.219)	(5.114)			(0.393)	(17.15)
Observations	387	387	434	434	332	332	434	434
R^2	0.839	0.765	0.727	0.665	0.871	0.756	0.758	0.652
Province FE	YES	YES	YES	YES	YES	YES	YES	YES
Tester FE	YES	YES	NO	NO	YES	YES	NO	NO
F-test	112.1	50.50	103.7	47.41	170.6	73.92	157	76.56
P-value	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001

Notes: Dependent variables are expressed in *z*-scores. *Z*-scores are obtained by non-parametrically standardizing within age (in month) groups. Standard errors in (parentheses) are computed with the cluster-correlated Huber-White estimator and adjusted for clustering at the village level and stratification at the township level.

Source: Author calculations.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Taken together, the linear coefficients of the impacts per home visits on child skills and health are larger during the first year than during the second year. We find evidence of decreasing returns on investment in skill development when 48 home visits are delivered over a two-year period. Moreover, our results indicate that investment in health is not productive during the second year. We argue that these findings indicate that if 48 bi-weekly home visits are delivered to children that are 6 to 18 months old at baseline over a period of two years, the optimal program duration or number of home visits is exceeded. Reducing program duration or the frequency of home visits can lead to gains in cost-effectiveness. It is possible, that it is optimal to end the program when children reach the nursery school entry age (i.e., at age 3), while encouraging parents to continue investing in their children after program completion (e.g., via enrolment in a good nursery school). Finally, we find no evidence that delivering information on child health during the second year of the intervention is effective.

REFERENCES

Angrist, Joshua D. 2014. “The Perils of Peer Effects.” *Labour Economics* 30 (October): 98–108. https://doi.org/10.1016/j.labeco.2014.05.008.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. “Generalized Random Forests.” *Annals of Statistics* 47 (2): 1179–1203. https://doi.org/10.1214/18-AOS1709.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. 2016. “Inference in High-Dimensional Panel Models With an Application to Gun Control.” *Journal of Business and Economic Statistics* 34 (4): 590–605. https://doi.org/10.1080/07350015.2015.1102733.

Sokal, R.R., and C.D. Michener. 1958. “A Statistical Method for Evaluating Systematic Relationships – ScienceOpen.” *University of Kansas Science Bulletin* 28: 1409–38.

Ward, Joe H. 1963. “Hierarchical Grouping to Optimize an Objective Function.” *Journal of the American Statistical Association* 58 (301): 236–44. https://doi.org/10.1080/01621459.1963.10500845.

Zubin, J. 1938. “A Technique for Measuring Like-Mindedness.” *Journal of Abnormal and Social Psychology* 33 (4): 508–16. https://doi.org/10.1037/h0055441.