I Z A Institute of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15132

The Impacts of a Prototypical Home Visiting Program on Child Skills

Jin Zhou
James Heckman
Bei Liu
Mai Lu

MARCH 2022

I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15132

The Impacts of a Prototypical Home Visiting Program on Child Skills

Jin Zhou
University of Chicago

James Heckman
University of Chicago and IZA

Bei Liu
China Development Research Foundation

Mai Lu
China Development Research Foundation

MARCH 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Impacts of a Prototypical Home Visiting Program on Child Skills* **

This paper uses random assignment to estimate the causal impacts on child skills of a widely emulated early childhood home visiting program. We show the feasibility of replicating it at scale. We estimate vectors of latent skills for individual children and compare treatments and controls. The program substantially improves child language and cognitive, fine motor, and social-emotional skills. We go beyond reporting treatment effects as unweighted item scores. We determine whether the program affects the latent skills generating correct answers to lists of test items and how the program affects the mapping from skills to item scores. Enhancements in latent skills explain most of the conventional treatment effects for language and cognition. The program operates primarily by improving skills and not by improving how effectively skills are used. The program barely changes the map from latent skills to item test scores.

JEL Classification: J13, Z18

Keywords: experiment, scaling, mechanisms, home visiting programs, measurement

Corresponding author:
Jin Zhou
Center for the Economics of Human Development
University of Chicago
1126 East 59th Street
Chicago, IL 60637
USA
E-mail: jinzhou@uchicago.edu

* CEHD acknowledges support from the Institute for New Economic Thinking, and the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number R37HD065072. The program has been registered at AEA with registry number AEARCTR-0007119. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the funders or the official views of the National Institutes of Health. CDRF acknowledges support from the UBS Optimus Foundation and the Dunhe Foundation. The authors wish to thank Susan Chang, Sally Grantham-McGregor, Sylvi Kuperman, Carey Cheng, Rebecca Myerson, Chunni Zhang, and Yike Wang for their efforts on program design, implementation, and data cleaning support. Erlfang Tsai and Fuyao Wang provided highly competent research assistance. CDRF thanks Mary Young, Fan Bu, Peng Liu, Lijia Shi, Bojiao Liang, and Yi Qie for their essential and valuable fieldwork support. We are grateful to the participants and their families for their continued participation in this research project. http://cehd.uchicago.edu/china-reach_home-visiting_appendix is a website for this paper with supplementary material.

** An earlier version of this paper is available as IZA DP No. 13346.

1 Introduction

A growing body of research establishes the effectiveness of home visiting programs targeted to the early years in developing the skills of disadvantaged children. Small-scale home visiting programs have been shown to be effective (see, e.g., Howard and Brooks-Gunn, 2009; HomVEE, 2020; Grantham-McGregor and Smith, 2016). They are relatively low cost compared to many other early childhood programs. They place minimal demands on the training required of the visitors and on the infrastructure needed to support them. Visitors have levels of education comparable to those of the caregivers visited. The Jamaica Reach Up and Learn program, established over 30 years ago, is a successful home visiting program emulated around the world (Grantham-McGregor and Smith, 2016).

This paper studies a large-scale replication of the original Jamaica program, China REACH, in a poor region of Western China (1500+ participants compared to the 100+ participants in the original Jamaica study). The program is evaluated by a randomized control trial, as was the original Jamaica program. Our evidence suggests that the program can be successfully implemented at scale.

The China REACH program has much richer data than the original Jamaica program, in part because the same group of scholars designed both projects and incorporated their lessons learned from Jamaica into the China version. We show that it has a strong impact on language and cognitive skills, fine motor skills, and social-emotional skills, but the impacts are not uniform across baseline distributions. Positive impacts on skills are strongest for children with absent mothers.

In securing these results, we depart from conventional practice and adjust for task difficulty levels across the multiple items used to assess skills. We thus avoid the unjustified but widely followed approach in the literature of reporting unweighted counts of performances on tasks that vary in difficulty. Our adjustments produce more plausible estimated treatment effects. We decompose estimated treatment effects into improvements in latent skills and improvements in the ability to use skills. Treatment effects primarily arise from boosts in skills.

This paper proceeds as follows. Section 2 describes the program. It is a scaled and enhanced version of the original Jamaica program. Section 3 presents an array of conventional experimental treatment effects and documents heterogeneity in program impacts. Furthermore, we estimate a nonlinear factor model with individual-level latent skills and determine the impact of treatment on the skills that generate item scores. Section 4 examines the sources of the estimated treatment effects. We examine the extent to which the program affects the inputs into the functions mapping skills to performance on tasks and the extent to which it shifts the productivity of a fixed stock of latent skills. Section 5 compares outcomes from the China program with those from the parent Jamaica program with follow-up through age 30. China REACH is on track to replicate Jamaica's long-term improvement of education and labor market outcomes. Section 6 summarizes our findings.

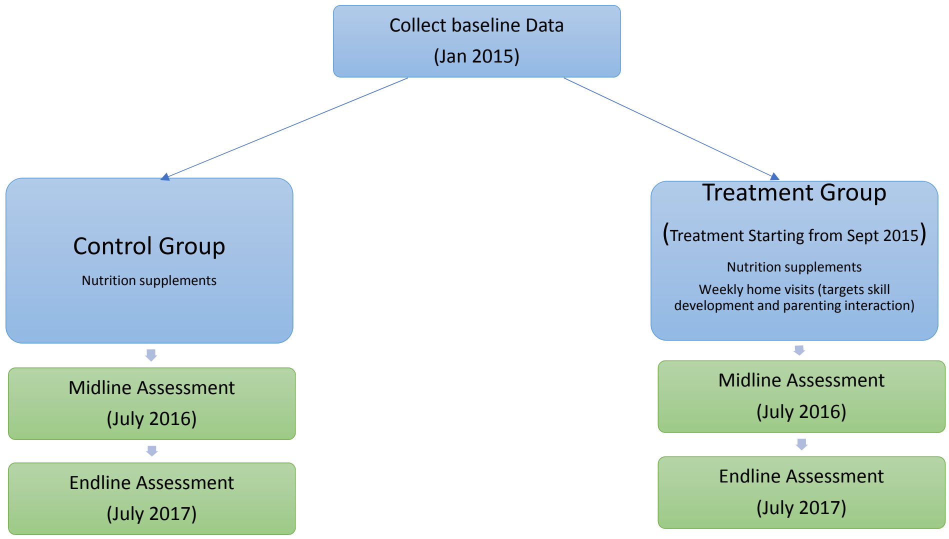
2 China REACH

The ongoing China Rural Education and Child Health (China REACH) project was launched in 2015 in response to a growing focus on, and call for, evidence-based pilot-to-policy analyses by China's State Council. It is a large-scale randomized control trial (RCT) designed to evaluate the impacts of a low-cost home visit delivery model for disadvantaged families. It is based on a successful Jamaican pilot (see Grantham-McGregor and Smith, 2016; Gertler, Heckman, Pinto, Zanolini, Vermeersch, Walker, Chang, and Grantham-McGregor, 2014). The program aims to improve the health and cognition of children by enhancing their engagement with caregivers and the larger community.

The program was conducted in Huachi County in Gansu Province, one of the poorest areas in China. The county has 15 townships, including 111 administrative villages. It is 85% mountainous with a population of 132,000, of whom 114,600 have rural hukou.[1] Figure 1 shows that the program we study was launched in January 2015 and that home visits started in September 2015. For details on program implementation, see Appendix A.

[1]Hukou is a type of household registration system in China that defines and limits mobility within China. There are agricultural and non-agricultural types of hukou.

Figure 1: The Timeline of China REACH (Huachi) Program



2.1 The Intervention Implemented

The program trains home visitors who have educational attainments at the level of the mothers visited. In rural China, it is easily replicated because the potential supply of home visitors is large. The program encourages child caregivers to interact with their children in developmentally appropriate ways. Lizzeri and Siniscalchi (2008) develop a model of child development that features parent-child interactions as important determinants of good parenting. Appendix B documents the home visiting protocols used.

Local implementation of the China REACH project is conducted by a county project coordinator, assisted by 24 township supervisors and 91 home visitors.[2] The coordinator prepares countywide training to oversee the township supervisors. The county project coordinator and township supervisors randomly attend home visits for spot checks to observe and review the home visitors' work.

The supervisors support and manage home visitors. They make sure that the home visitors prepare for weekly visits, review the content of past visits, plan ac-

[2]Townships are geographic partitions of the entire county. On average, each home visitor is in charge of eight households' home visits.

tivities for future visits, and organize weekly meetings with the home visitors to improve and reflect on the home visiting program and experience. Township supervisors visit each household with the home visitor once a month and record observations on the caregiver, child, and home visitor and their interactions.

The visitors engage with households weekly and provide one hour of parenting or caregiving guidance and support based on the Jamaica program protocols.[3] During each home visit, the home visitor records information about parental engagement (e.g., who worked with the child during the visit, whether the home visitor taught parents relevant tasks if the child could not participate in the home visit, and who played with the child after the visit and with what frequency) and child performance (e.g., tasks taught in the last week and new tasks taught in the current week). Appendix B.3 documents the content of the China REACH curriculum, the content of each weekly visit, and the assessment instruments used each week. The curriculum includes more than 200 tasks related to language and cognitive skill development and has about 70 fine motor tasks and 20 tasks targeting gross motor skill development.

2.1.1 Design of the Randomized Control Trial

Randomization is based on a village- (cluster-) level matched-pair design. Bai (2019) shows that this design is optimal for minimizing the mean-squared error of estimates of average treatment effects. Implementation is in three steps. We first examine the entire universe of eligible villages in Huachi county. We next use household surveys and village-level administrative data to assess the similarities of villages using a Mahalanobis metric of resident and village characteristics.[4]

[3]The protocols are based on those used by the Jamaica program but adapted to Chinese culture (e.g., by changing the songs to popular Chinese songs and adding backgrounds familiar to Chinese people). The protocol for children younger than 18 months focuses on motor and language skill training. For those older than 18 months, the protocol adds more cognitive skill content (e.g., classification, pairing, and picture puzzles).

[4]The pre-treatment village-level covariates used for the matching village pairs include: (1) the "closeness with children" scores on the Home Observation for Measurement of the Environment Inventory (HOME IT) scale; (2) the language skill scores on the HOME IT scale; (3) the learning materials score on the HOME IT scale; (4) the take-up rate of a nutrition supplement program in the village; (5) the compliance rate for a countywide nutrition program in the village; (6) the percentage of left-behind children in the children sample; (7) the per capita net income in the village; (8) the average years of schooling in the village; (9) the percentage of caregivers intending to participate in the parenting intervention program; and (10) the percentage of families intending to bring the child when migrating to urban areas.

To minimize the Mahalanobis metric in each pair, we sort the villages by metric scores and pair the closest ones using the nonparametric belief propagation (nbp) matching method.[5]

After matching village pairs, we randomly select one village within the pair into the treatment group and the other village into the control group.[6] Figure A.2 in the Appendix indicates the location of the paired villages in Huachi county. The design closely matches the characteristics of the villages in the pairs.[7] Village-level treatment effects include within-village spillovers. Villages are used only once, as either treatments or controls.

3 Estimated Treatment Effects

The China REACH intervention aims to promote multiple skills (e.g., motor, language, cognitive, and social-emotional skills). Table 1 displays our measures of skill. The Denver II test provides detailed child development assessment task measures.[8,9,10]

[5] Lu, Greevy, Xu, and Beck (2011).

[6] In total, there are 55 matched pairs, which means there are 55 villages in both the treatment and control groups.

[7] Appendix C documents baseline comparisons.

[8] The Denver II test is designed for clinicians, teachers, or early childhood professionals monitoring the development of infants and preschool-age children. The test is primarily based on the examiner's actual observations rather than a parental report. It is an inventory of 125 tasks, including four types of skill measures: personal-social (caring for personal needs and getting along with people), fine motor–adaptive (hand-eye coordination, manipulation of small objects, and problem-solving), language (hearing, understanding, and using language), and gross motor (sitting, walking, jumping, and overall large muscle movement).

[9] Appendix D gives both the English and Chinese versions of the Denver II test measure tables.

[10] The Bayley III test converts composite scores into scaled scores based on age, which are more useful in clinical practice. However, it is also possible to achieve the same goal by using itemized Denver II test measures. The Bayley III test targets infants and children between 1 and 42 months of age and includes both the examiner's observations (cognitive, motor, and language skills) and the parents' questionnaires (social-emotional and adaptive behavior skills). Ryu and Sim (2019) report that the Denver test is more accurate than the Bayley test in detecting the delay of language development.

Table 1: China REACH Home Visiting Program Skill Content

Skill Category	Definition
Fine Motor	The skill of finger movements, such as grasping, releasing and stitching, drawing, and writing.
Gross Motor	A wide range of body muscle movements, such as walking, running, throwing, and kicking.
Cognitive	The skill of learning, which includes logic, problem-solving, memory, and attention.
Language	Vocalization, gestures, and speaking coherent words.
Social-Emotional	Express and control emotions and communicate in a developmentally appropriate way.

This section reports conventional estimates of the home visiting intervention's average treatment effects on unweighted sums of item scores within each category. Item scores are binary indicators of knowledge of a task. We use robust statistical methods to adjust for missing data and allow disturbances within villages to be correlated (Cameron, Gelbach, and Miller, 2008).

Using the proportion of items correctly answered as an outcome, which is standard practice, assumes that the test difficulty levels are the same for each task. In practice, there is substantial variation in the task difficulty levels in the Denver II test we use. We address this problem using a nonlinear measurement model that accounts for item difficulty (van der Linden, 2016) and recover *individual* latent skills that generate item responses. We identify both experimentally induced improvements in latent skills and improvements in utilization of skills to answer individual test questions.

3.1 County-Level Average Treatment Effects

We now define the treatment effects we report. To facilitate exposition, it is helpful to define some notation. The universe of villages is $\{1,\ldots,V\}$. Villages are paired by a matching rule $m(v) : v \to v'$ where v' is the closest match to v in terms of a vector of mean pre-treatment covariates $\bar{\boldsymbol{Z}}(v)$. Proximity is calibrated by a Mahalanobis metric:

$$v' = \underset{\{1,\ldots,V\}\setminus\{v'\}}{\operatorname{argmin}} \left(\bar{\boldsymbol{Z}}(v) - \bar{\boldsymbol{Z}}(v')\right)' \sum \left(\bar{\boldsymbol{Z}}(v) - \bar{\boldsymbol{Z}}(v')\right)$$

where $\sum$ is the covariance matrix of $\boldsymbol{Z}$ computed over all villages. A coin is tossed to determine which village of a (v, v') pair receives treatment. No village is used twice.

Let $D_v = 1$ if v is selected into treatment. All individuals i are assigned to some village. $D_{v(i)}$ is the assigned treatment status of i in v, $D_{v(i)} \in \{0, 1\}$. Each village has I_v eligible inhabitants.

We first report average treatment effects for standardized scores estimated from the following empirical model:

$$Y_{iv}^{m} = \boldsymbol{\beta_0} + D_{v(i)}\boldsymbol{\beta_1^m} + \boldsymbol{Z_i'}\boldsymbol{\beta_2^m} + \sum_{p=1}^{P} 1\{i \in p\}\boldsymbol{\beta_p^m} + \varepsilon_{iv}^{m} \tag{1}$$

where Y_{iv}^{m} are the standardized scores for outcome m for child i in village v, $D_{v(i)}$ is a dummy variable indicating the treatment status of village v in which child i lives, and $\boldsymbol{Z_i}$ are the pre-treatment covariates. $1\{i \in p\}$ is an indicator of whether the child i lives in the village pair p. $Y_{iv}^{m} = D_{v(i)}Y_{iv}^{m}(1) + (1 - D_{v(i)})Y_{iv}^{m}(0)$, where $Y_{iv}^{m}(d)$ denotes the vector of outcomes fixing treatment status d. The treatment assignment design implies that

$$\left(Y_{iv}^{m}(0), Y_{iv}^{m}(1)\right) \perp\!\!\!\perp D_{v(i)} \mid \boldsymbol{Z_i}. \tag{2}$$

Treatment is at the village level. The idiosyncratic shock term ε_{iv}^{m} for child i can be arbitrarily correlated with $\varepsilon_{i'v}^{m}$ for any other child $i' \neq i$ in the same village v. Idiosyncratic shocks are assumed to be independent across villages; i.e., $\varepsilon_{iv}^{m} \perp\!\!\!\perp \varepsilon_{kv'}^{m}$ for $\forall i \in v$ and $\forall k \in v', v \neq v'$. Residual plots displayed in Appendix E verify the assumption of independence of residuals across villages. The $N \times N$ covariance matrix $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \boldsymbol{\Omega}$ with V number of villages is block diagonal: $\boldsymbol{\Omega}_{vv'} = 0$; all $v \neq v'$.[11]

As the number of observations in each cluster gets large, and as the number of clusters gets large, the OLS estimator of the parameters of (1) is consistent, provided that the ratio of clusters to observations in the cluster converges to a constant. This is true if $\boldsymbol{\beta_1^m}$ is constant across people.

Define the full array of right-hand side variables in (1) by $\boldsymbol{X_{iv}}$. The standard cluster-robust variance estimator (CRVE), $(\boldsymbol{X'X})^{-1}(\sum_{v=1}^{V} \boldsymbol{X_v'}\hat{\boldsymbol{\Omega}}_v\boldsymbol{X_v})(\boldsymbol{X'X})^{-1}$, is bi-

[11] $\boldsymbol{X}_v$ indicates $\boldsymbol{X}$ in the vth cluster, and $E(\boldsymbol{\epsilon}_v) = 0$, $E(\boldsymbol{\epsilon}_v\boldsymbol{\epsilon}_v') = \boldsymbol{\Omega}_v$. $\boldsymbol{X}$ includes the treatment status, pre-treatment covariates, and indicators of the matched pair.

ased when $\hat{\Omega}_v$ is estimated using the OLS residuals $\hat{\epsilon}_v$: $E(\hat{\boldsymbol{\epsilon}}_v\hat{\boldsymbol{\epsilon}}_v')$.[12] The bias depends on the form of $\boldsymbol{\Omega}_v$. Cameron, Gelbach, and Miller (2008) discuss this problem and show that the wild cluster bootstrap performs well in making cluster-robust inferences. Details of the wild bootstrap procedures we use are presented in Appendix F.[13]

In our sample, over 98% of eligible children in the treated villages receive home visits. Still, about 15% of children from both the control and treatment groups miss the annual child development assessment. To obtain consistent estimates of population average treatment effects, we use inverse probability weighting (Tsiatis, 2006).[14,15]

Table 2 presents the treatment effects for each skill category using standardized outcome measures.[16,17] Using different statistical models, columns (1), (2), and (3) use all available data samples, and columns (4) and (5) only use samples of children who are under 2 years of age in September 2015 when the program started. The younger treated children have at least one year of exposure to the intervention.[18]

The first row in Table 2 shows that the children in the treatment group are, on average, more likely to have higher language and cognitive skills.[19] In the first

[12] $\hat{\boldsymbol{\epsilon}}_v$ are the OLS residuals.

[13] Because we have 55 clusters, recent concerns about the wild bootstrap do not apply. See Canay, Santos, and Shaikh (2019).

[14] Maasoumi and Wang (2019) provide robust inference using the IPW method to trim out low-probability observations. In our paper, only three observations' propensity scores (of being non-missing) are lower than 0.1. Therefore, we do not need to trim the data and can avoid the inconsistency problem.

[15] Appendix G documents the data attrition problem and how we construct the probability of missing data. To avoid redundancy, we include inverse probabilities in all estimations in the paper.

[16] Only 140 children took the Denver test at the baseline. We estimate the same model for the children with baseline information and do not find significant differences in Denver test scores between the control and treatment groups. The details about this balancing test are presented in Appendix C.

[17] There is no population-level reference for the Denver test in China. We use the control group as the reference group: we estimate Denver test performance by monthly age and then use the mean and the variance to standardize the test scores at each monthly age group for both the treatment and control groups.

[18] There are two reasons for restricting the sample. (1) As claimed, we want the children in the treatment group to have substantial exposure to the intervention. Many older children participate for shorter periods of time. (2) We have more older children in the control group than in the treatment group because the field team did not update the name list in the treatment group after September 2015.

[19] We combine these categories to obtain a number of item scores comparable to the number we have for the other categories.

row, we see that at midline (about nine months into the intervention) the language and cognitive skills of the children in the treatment group are about 0.7 standard deviations higher than those of the children in the control group. At the end of the intervention, effect sizes for treatment effects on language and cognitive skills are greater than 1. The intervention significantly improves treated children's language and cognitive skills. The magnitudes of the age-adjusted treatment effects increase when the children in the treatment group have earlier and hence longer exposure to home visitors (see columns (4) and (5)). This is consistent with dynamic complementarity.

The intervention significantly improves social-emotional skills at midline and fine motor skills at the end of the intervention but produces no significant improvement in gross motor skills. This finding is consistent with the design of the curriculum, which focuses primarily on language and cognitive skill development.[20,21]

Tables 3–4 display treatment effects by gender. An interesting finding, consistent with recurrent findings in the literature (Elango, García, Heckman, and Hojman, 2016), is that the intervention improves boys' language and cognitive skills much more than those of girls. At midline, the treatment effect size is 0.4 for girls and 0.9 for boys. At the end of the intervention, the effect size is about 0.9 for girls and 1.1 for boys. One reason for this is that girls are, on average, relatively more developed than boys at the same age in early childhood. The girls in the treatment group also have better performance in social-emotional skills.[22]

[20] Heckman and Zhou (2021) document the intervention curriculum.

[21] Results are comparable when we use raw rather than standardized scores. These are reported in Appendix E.

[22] This result is also found in the evaluation of the Perry Preschool program (Heckman and Karapakula, 2019) and the Abecedarian preschool program (García, Heckman, and Ziff, 2018).

Table 2: Treatment Effects on Standardized Denver Scores

	(1)	(2)	(3)	(4)	(5)
	All	All	All	Children $\leq$ 2 Yrs at Enrollment	Children $\leq$ 2 Yrs at Enrollment
			Midline		
Language and Cognitive	0.589***	0.631***	0.714***	0.674***	0.741***
	[0.234, 0.965]	[0.237, 1.036]	[0.319, 1.093]	[0.279, 1.067]	[0.350, 1.144]
Fine Motor	0.334	0.559	0.633*	0.629*	0.703*
	[-0.140, 0.787]	[-0.032, 1.174]	[0.003, 1.313]	[0.023, 1.324]	[0.057, 1.375]
Social-Emotional	0.690**	0.865***	0.879***	0.624***	0.620***
	[0.260, 1.117]	[0.421, 1.312]	[0.467, 1.289]	[0.129, 1.118]	[0.204, 1.067]
Gross Motor	-0.051	-0.004	-0.015	0.054	0.010
	[-0.598, 0.478]	[-0.564, 0.577]	[-0.567, 0.554]	[-0.514, 0.640]	[-0.559, 0.584]
			Endline		
Language and Cognitive	0.979***	0.914***	1.036***	1.016***	1.113***
	[0.585, 1.402]	[0.495, 1.347]	[0.644, 1.458]	[0.637, 1.408]	[0.723, 1.510]
Fine Motor	0.585**	0.574**	0.676***	0.561**	0.645**
	[0.006, 0.956]	[0.067, 1.091]	[0.180, 1.170]	[0.030, 1.095]	[0.139, 1.158]
Social-Emotional	-0.201	-0.276	-0.222	-0.167	-0.115
	[-0.596, 0.202]	[-0.688, 0.123]	[-0.636, 0.194]	[-0.553, 0.215]	[-0.491, 0.275]
Gross Motor	0.067	0.125	0.173	0.155	0.219
	[-0.479, 0.632]	[-0.392, 0.645]	[-0.322, 0.668]	[-0.406, 0.732]	[-0.294, 0.775]
Pre-treatment Covariates	No	No	Yes	No	Yes
IPW	No	Yes	Yes	Yes	Yes

Notes: 1. The 95% confidence intervals in brackets are constructed using the wild bootstrap clustered at the village level.
2. The mean and variance for the standardized score are estimated from the pooled sample of the control group children.
3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
4. The negative treatment effects for social-emotional ability vanish after we adjust for item difficulty.
5. The columns with the label “All” include all the observations, and the columns with the label “Children $\leq$ 2 Yrs at Enrollment” restrict the sample to the children who were under 2 years old when they enrolled in the program.

Table 3: Treatment Effects on Standardized Denver Scores

(Female)

	(1)	(2)	(3)	(4)	(5)
	All	All	All	Children $\leq$ 2 Yrs at Enrollment	Children $\leq$ 2 Yrs at Enrollment
			Midline		
Language and Cognitive	0.410	0.417	0.445	0.511**	0.534**
	[-0.076, 0.869]	[-0.035, 0.884]	[-0.014, 0.910]	[0.040, 0.991]	[0.080, 0.990]
Fine Motor	0.400	0.399	0.335	0.512	0.544
	[-0.252, 1.049]	[-0.271, 1.065]	[-0.269, 1.211]	[-0.088, 1.142]	[-0.082, 1.189]
Social-Emotional	1.020***	1.068***	1.114***	0.912**	0.938***
	[0.445, 1.614]	[0.520, 1.614]	[0.681, 1.550]	[0.272, 1.541]	[0.400, 1.431]
Gross Motor	0.117	0.063	0.058	0.085	0.019
	[-0.487, 0.751]	[-0.565, 0.665]	[-0.532, 0.675]	[-0.514, 0.725]	[-0.605, 0.652]
			Endline		
Language and Cognitive	0.852**	0.895**	0.950**	0.865**	0.893**
	[0.077, 1.596]	[0.159, 1.612]	[0.213, 1.675]	[0.122, 1.590]	[0.177, 1.598]
Fine Motor	0.804**	0.815**	0.866**	0.836**	0.855**
	[0.111, 1.500]	[0.088, 1.553]	[0.189, 1.574]	[0.110, 1.554]	[0.117, 1.579]
Social-Emotional	-0.264	-0.298	-0.309	-0.264	-0.291
	[-0.806, 0.254]	[-0.805, 0.267]	[-0.775, 0.160]	[-0.859, 0.342]	[-0.820, 0.206]
Gross Motor	0.188	0.246	0.257	0.460	0.445
	[-0.737, 1.091]	[-0.668, 1.094]	[-0.582, 1.080]	[-0.410, 1.308]	[-0.417, 1.326]
Pre-treatment Covariates	No	No	Yes	No	Yes
IPW	No	Yes	Yes	Yes	Yes

Notes: 1. The 95% confidence intervals in brackets are constructed using the wild bootstrap clustered at the village level.
2. The mean and variance for the standardized score are estimated from the pooled sample of the control group children.
3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
4. The negative treatment effects for social-emotional ability vanish after we adjust for item difficulty.
5. The columns with the label "All" include all the observations, and the columns with the label "Children $\leq$ 2 Yrs at Enrollment" restrict the sample to the children who were under 2 years old when they enrolled in the program.

Table 4: Treatment Effects on Standardized Denver Scores

(Male)

	(1)	(2)	(3)	(4)	(5)
	All	All	All	Children $\leq$ 2 Yrs at Enrollment	Children $\leq$ 2 Yrs at Enrollment
			Midline		
Language and Cognitive	0.747***	0.852***	0.938***	0.896***	0.911***
	[0.236, 1.257]	[0.261, 1.462]	[0.389, 1.499]	[0.345, 1.460]	[0.329, 1.501]
Fine Motor	0.395	0.674	0.716	0.730	0.771
	[-0.108, 0.908]	[-0.083, 1.532]	[-0.099, 1.598]	[-0.028, 1.577]	[-0.070, 1.747]
Social-Emotional	0.436	0.589*	0.549**	0.395	0.280
	[-0.115, 0.989]	[0.028, 1.140]	[0.047, 1.054]	[-0.178, 0.946]	[-0.272, 0.842]
Gross Motor	-0.066	0.079	-0.041	0.152	-0.021
	[-0.798, 0.661]	[-0.728, 0.900]	[-0.700, 0.639]	[-0.634, 0.963]	[-0.682, 0.659]
			Endline		
Language and Cognitive	1.050***	0.797**	0.950***	1.000***	1.111***
	[0.514, 1.560]	[0.205, 1.436]	[0.448, 1.497]	[0.468, 1.513]	[0.625, 1.626]
Fine Motor	0.460	0.388	0.462	0.346	0.388
	[-0.212, 1.117]	[-0.314, 1.108]	[-0.206, 1.144]	[-0.374, 1.042]	[-0.355, 1.124]
Social-Emotional	-0.139	-0.306	-0.256	-0.157	-0.169
	[-0.643, 0.390]	[-0.895, 0.305]	[-0.829, 0.326]	[-0.654, 0.351]	[-0.701, 0.400]
Gross Motor	-0.059	-0.071	-0.048	-0.169	-0.138
	[-0.528, 0.424]	[-0.543, 0.407]	[-0.510, 0.419]	[-0.663, 0.332]	[-0.629, 0.359]
Pre-treatment Covariates	No	No	Yes	No	Yes
IPW	No	Yes	Yes	Yes	Yes

Notes: 1. The 95% confidence intervals in brackets are constructed using the wild bootstrap clustered at the village level.
2. The mean and variance for the standardized score are estimated from the pooled sample of the control group children.
3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
4. The negative treatment effects for social-emotional skills vanish after we adjust for item difficulty.
5. The columns with the label "All" include all the observations, and the columns with the label "Children $\leq$ 2 Yrs at Enrollment" restrict the sample to the children who were under 2 years old when they enrolled in the program.

Appendix H presents an analysis of the impacts on child skills of interactions between the home visitor and the caregiver, and the home visitor and the child, as well as variables capturing home visitor teaching ability.[23] The only strong pattern that emerges is that good caregiver–home visitor interactions promote language and cognitive skills.[24]

3.2 Adjusting for Item Difficulty and Estimating the Effect of Treatment on Latent Skills

The previous analysis shows that treatment boosts outcomes on unweighted item aggregates. Aggregates so formed, while traditional, are problematic unless the difficulty is the same across tasks, which is not true by the design of the assessments.

To address this issue, we take advantage of the multi-item nature of our data and estimate a nonlinear factor model with individual-level latent skills.[25] We follow standard methods in psychometrics and introduce and estimate difficulty parameters across items (van der Linden, 2016). We also estimate individual-level latent skills. We use our estimates to determine the impact of treatment on the skills that generate item scores. We also estimate how much the intervention shifts the map between skills and item scores (i.e., whether treated children better utilize existing skills).

3.2.1 Items and Skills

The outcomes we study are children's performances on individual tasks measured by performance on items on a test. There are N_{J_k} tasks for each of the K distinct skills. Tasks are skill-specific (e.g., motor, cognitive, reading, etc). Performance on the tasks is assumed to be generated by latent skills $\boldsymbol{\theta}$. We use N_J to denote the total number of items for all skills (i.e., $N_J = \sum_{k=1}^{K} N_{J_k}$). We assume

[23] Measures of interactions are recorded monthly. The measures used for the midline regression are means taken over monthly measures through midline. The measures used for the endline regression are means of the measures over the entire intervention.

[24] Table H.2 in Appendix H shows considerable dispersion in these measures, so the weak estimates of the interaction effects are not due to inadequate sample variance.

[25] In the data, we have more than 70 items per skill per individual on which to measure task performance on the Denver test.

that a common technology mapping skills to test scores operates in all villages. We thus drop the v-specific notation. Let $Y_i^{j_k}(d)$ be a binary-valued outcome variable indicating mastery of task j for skill type k by person i. Performance is generated by a latent outcome for task item j for a person with treatment status $d \in \{0, 1\}$. Let $\boldsymbol{\theta_i^d}$ be a K-dimensional vector of latent skills for person with treatment status d. $\boldsymbol{X_i}$ is a vector of baseline covariates. Write the mapping from latent skills $\boldsymbol{\theta_i^d}$ to the determinants of outcome on task j as

$$\tilde{Y}_i^{j_k}(d) = \boldsymbol{X_i'}\boldsymbol{\beta}^{\boldsymbol{j_k,d}} + \delta^{j_k} + (\boldsymbol{\theta_i^d})'\boldsymbol{\alpha}^{\boldsymbol{j_k,d}} + \varepsilon_i^{j_k}, \quad j = 1, \ldots, N_{J_k}; k = 1, \ldots, K. \tag{3}$$

$$Y_i^{j_k}(d) = \begin{cases} 1 & \tilde{Y}_i^{j_k}(d) \geq 0 \\ 0 & \tilde{Y}_i^{j_k}(d) < 0 \end{cases}$$

where $\boldsymbol{\alpha}^{\boldsymbol{j_k,d}}$ is a K-dimensional vector of factor loadings; δ^{j_k} is a task difficulty parameter for the task item j_k; and the coefficients $\boldsymbol{\beta}^{\boldsymbol{j_k,d}}$ and $\boldsymbol{\alpha}^{\boldsymbol{j_k,d}}$ can depend on treatment, the skills modeled, and even the item studied, where items are common across people. In estimation, we impose $\boldsymbol{\beta}^{\boldsymbol{j_k,d}} = \boldsymbol{\beta}^{\boldsymbol{j_k',d}} = \boldsymbol{\beta}^{\boldsymbol{k,d}}$, $\forall j_k$ and j_k'; i.e., coefficients are common across items within a skill.

This model interprets the intervention as shaping skills that affect performance on tasks. The intervention may also enhance the productivity of any given skill in performing a task; i.e., the intervention shifts $\boldsymbol{\alpha}^{\boldsymbol{j_k,d}}$. The object $(\boldsymbol{\theta_i^d})'\boldsymbol{\alpha}^{\boldsymbol{j_k,d}}$ is a bundle of effective skills for outcome j_k from intervention $D = d$ arising from either source.

Under suitable normalizations, we can identify the *individual*-level latent skill factors $\boldsymbol{\theta}_i^{\boldsymbol{d}}$ and not just the distribution of the latent skill factors, as in traditional psychometric models (see, e.g., van der Linden, 2016). We assume that $\varepsilon_i^{j_k}$ is unit normal, independent of the other right-hand side variables. This data has a panel-like structure over items. It can be fit using a probit model with latent skills. We estimate the parameters of observed covariates, the latent factors, and the effects of latent skill factors on outcomes. From the analysis of Wang (2020), it can be shown that estimators of the parameters of the model, including individual abilities, are consistent and asymptotically unbiased when the number of observations (sample participants) $N_I \to \infty$ and $N_J \to \infty$ but $\frac{N_I}{N_J}$ converges to a constant.[26] These con-

[26]Recall that in estimation, the number of items is allowed to vary depending on the actual test

ditions apply in our sample with large numbers of test items per person and large numbers of observations.

Factor models require normalizations if one seeks to isolate $\boldsymbol{\theta}^{d}$ from $\boldsymbol{\alpha}^{j_k,d}$. Since $\boldsymbol{\theta}_i^{d\prime}\boldsymbol{\alpha}^{j_k,d} = (\boldsymbol{\theta}_i^{d})'\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{\alpha}^{j_k,d}$, the factors and factor loadings are intrinsically arbitrary unless a scale is somehow set. We can avoid such normalizations if we are content to measure the shifts in effective skills, $\boldsymbol{\theta}_i^{d\prime}\boldsymbol{\alpha}^{j_k,d}$. We can break this term apart using a normalization suggested by Anderson and Rubin (1956) and identify both the vector $\boldsymbol{\theta}_i^{d}$ and $\boldsymbol{\alpha}^{j_k,d}$. We report estimates for $\boldsymbol{\theta}_i^{d}$ and $\boldsymbol{\alpha}^{j_k,d}$ separately and also as a bundle of effective skills $(\boldsymbol{\theta}_i^{d})'\boldsymbol{\alpha}^{j_k,d}$.

Following traditions in the Rasch model literature (van der Linden, 2016), we assume that δ^{j_k} is a treatment-invariant task difficulty parameter intrinsic to the measurement system and independent of treatment status. This assures comparability of measurements across treatments and controls.

We have four different latent skill factors in our model, corresponding to social-emotional, language and cognitive, fine motor, and gross motor skills in the Denver II test $k \in \{1, \ldots, 4\}$. To interpret the factors, we assume that performance on K of N_J tasks ($K \leq N_J$) depends only on one factor. This specializes what Cunha, Heckman, and Schennach (2010) call the "dedicated factor case" to apply to only the first four items of each measurement. We thus generalize their analysis by requiring that only a subset of tasks are dedicated for any measurement of skills. We normalize the factor loading matrix so that the first K rows form an $\boldsymbol{I}_{\boldsymbol{K},\boldsymbol{K}}$ identity matrix. For the first $K = 4$ items of the measurements, we assume that they load on only one skill.[27] The remaining factor loading matrix for the vector of N_J outcomes is unrestricted. Dropping the d superscript to reduce notational clutter, we write the metric of loadings on the latent skills as $\boldsymbol{\alpha}'_{N_J \times K}$:

design.

[27]We select the washing and drying hands item, the imitate vertical line item, the combine words item, and the broad jump item to present social-emotional skills, fine motor skills, language and cognitive skills, and gross motor skills, respectively. Washing and drying hands is an important social skill in China due to its emphasis on hygiene and safe social environments.

$$
\boldsymbol{\alpha}'_{N_J \times K} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \alpha^{5,1} & \alpha^{5,2} & \alpha^{5,3} & \alpha^{5,4} \\ \vdots & \alpha^{6,2} & \dots & \dots \\ \alpha^{N_J,1} & \dots & \dots & \alpha^{N_J,4}. \end{bmatrix} \tag{4}
$$

We test and reject the "dedicated model" that assumes that in rows j_k of (4), for $j_k \geq 5$, $\alpha^{j_k,\ell,d} = 0$ except for one $\ell \in \{1, \dots, 4\}$. Table 5 reports this test. The assumption of a dedicated factor model fails in our sample.

Table 5: Test of Hypothesis for $j_k \geq 5$, $\alpha^{j_k,\ell,d} = 0$ except for one $\ell \in \{1, \dots, 4\}$

	Control		Treatment	
	$\chi^2(68)$	p-value	$\chi^2(68)$	p-value
Social-Emotional	463.247	0.000	1434.742	0.000
Fine Motor	494.200	0.000	1418.862	0.000
Language and Cognitive	1186.793	0.000	2108.501	0.000
Gross Motor	1570.322	0.000	1969.099	0.000

We report sensitivity analyses of our estimates using a variety of plausible normalizations in Appendix I. We find that the estimates of $\boldsymbol{\alpha}^{j_k,d}$ reported in the text are stable under a variety of different normalizations.[28] Our results are quantitatively robust. We use the estimation procedure proposed by Chen, Fernández-Val, and Weidner (2021) to estimate panel probit models with multiple latent skill factors.[29] The asymptotic justification for this approach for estimating individual-specific factors and population factor loadings is based on Wang (2020).

3.2.2 Estimates

Table 6 presents estimates of $\boldsymbol{\beta}^{k,d}$. There are no statistically significant differences between the treatment and control groups, although the point estimates for

[28]In Appendix I, we compare the distribution of the skill loadings under different normalizations. We find that the results are robust when we choose items within the median difficulty level range.

[29]Details regarding the method are presented in Appendix J.

males are substantially more negative for the treatment group. Figure 2 compares the distribution of the predicted combined language and cognitive task items from our model and the actual task items.[30] We also fit the data well with the other types of tasks.[31]

Table 6: Estimates of the Coefficients of the Observed Covariates

	Control Group	Treatment Group
Monthly Age	0.961	0.924
	[0.166, 1.987]	[0.161, 1.738]
Monthly Age^2	-0.009	-0.009
	[-0.025, 0.002]	[-0.0193, 0.002]
Male	0.356	-0.144
	[-1.081, 2.363]	[-1.178, 1.148]
Constant	-16.756	-15.571
	[-35.260, -2.727]	[-31.620, -2.457]
	$\chi^2(4) = 0.004$	$p = 0.999$

Notes: 1. The values presented in the brackets are 95% confidence intervals.
2. The confidence intervals are calculated by the paired cluster bootstrap at the village level.
3. We use the likelihood ratio test to examine whether the coefficients of two groups are the same or not. The test results show that we cannot reject the hypothesis that these coefficients are the same.

[30] We combine language and cognitive tasks into one category because of the paucity of cognitive test items in our Denver test.

[31] See Appendix K.

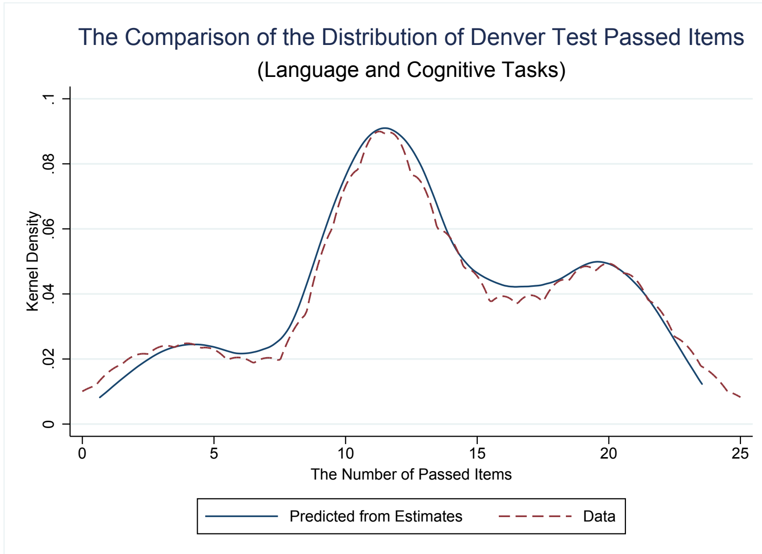


Figure 2: The Distribution of Denver Test Passed Items

Figure 3 shows the array of estimated difficulty level parameters δ^{j_k} for each task item. When the item difficulty level increases, the estimates become more negative. The estimates generally accord with the design of tests to increase the difficulty level with later items. The estimated difficulty level parameters δ^{j_k} provide information about whether the test is well designed. For example, the test for gross motor skills is not especially well designed: values of the difficulty level are flat around -1.8 and then quickly jump to -6 by the fifth item. This means that the children who took the test could correctly answer easy items but were likely to fail to answer all harder questions. Compared to gross motor skills task items, language and cognitive task items are better designed since the difficulty level rises smoothly across all items. The estimates of the social-emotional task items, however, do not accord with the intended assessment design.

Table 7: Treatment Effects on Mean of Latent Skill Factors

	Social-Emotional	Fine Motor	Language and Cognitive	Gross Motor
Treatment	0.395***	0.726***	0.753***	-0.095
	[0.208, 0.583]	[0.551, 0.899]	[0.459,1.051]	[-0.280, 0.089]

Notes: 1. The 95% confidence intervals in brackets are constructed by wild bootstrap clustered at the village level.
2. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

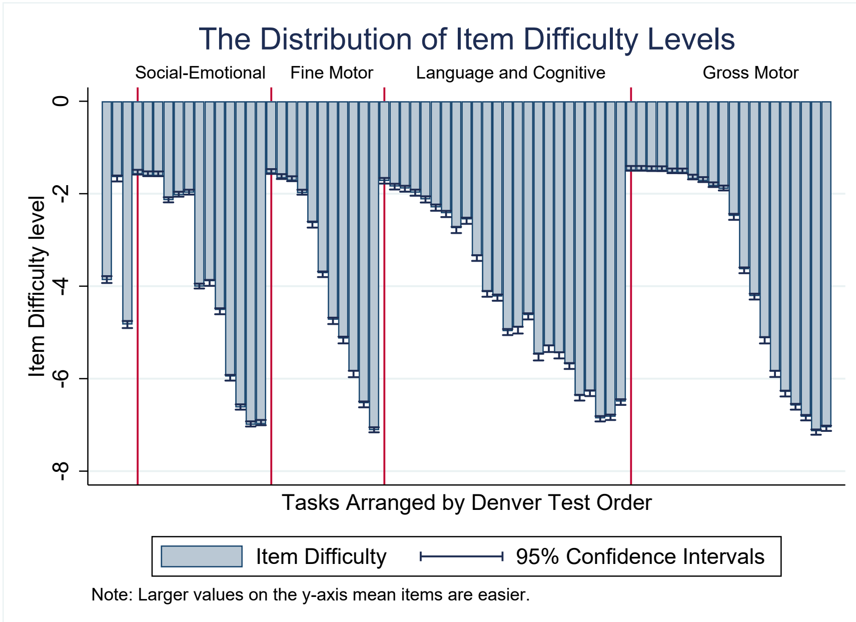


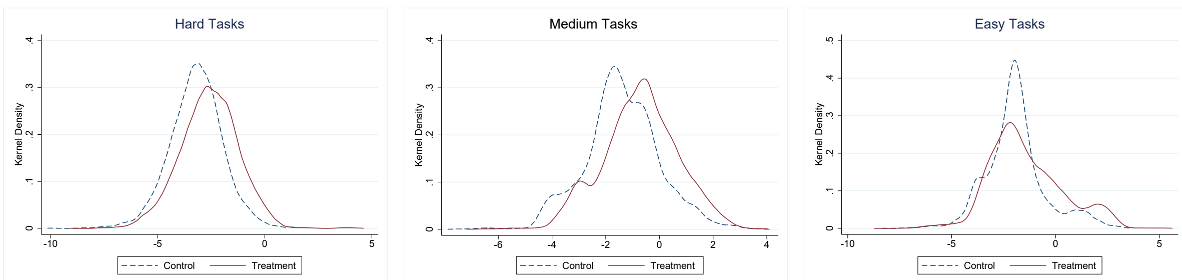
Figure 3: The Distribution of Denver Task Item Difficulty Levels

Table 8: The Correlation between Different Latent Skill Factors

	Social-Emotional	Fine Motor	Language and Cognitive	Gross Motor
Social-Emotional	1			
Fine Motor	0.428***	1		
Language and Cognitive	0.455***	0.207***	1	
Gross Motor	0.085***	0.156***	-0.102***	1

Notes: 1. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

An advantage of our approach is that we can estimate individual-level latent skill factors. First, Table 7 presents the treatment effects for the means of the four latent skill factors. Except for gross motor skills, the means of all other latent skill factors in the treatment group are significantly higher than those in the control group. When we compare treatment effects across different latent skills, we find that improvements in fine motor and language skills are at the same level but that there are no effects on gross motor skills. Table 8 shows that language and cognitive skills are negatively correlated with gross motor skills and positively correlated with social-emotional and fine motor skills.



(a) Hard Tasks (b) Medium Tasks (c) Easy Tasks

Figure 4: The Distributions of $\left[(\theta_i^d)'\alpha^{j_k,d}\right]^\dagger$

† There are 72 tasks ordered by estimated difficulty levels. Easy tasks are defined as those with difficulty parameters ranked between 1 and 24, medium tasks are those with difficulty parameters ranked between 25 and 48, and hard tasks are those with difficulty parameters ranked between 49 and 72.

Figure 4 plots the products of estimated skill factor loadings and the latent skill factors based on the Denver task difficulty levels.[32] The loadings for the treatment group are larger for the hard and medium tasks but smaller for the easy tasks, which indicates that the easier tasks are not helpful for detecting treatment effects on child skill development. The loadings have similar patterns across the treatment and control groups for other skills. Estimates of aggregates of loadings are precisely estimated, and for most tasks, we reject the hypothesis that $\alpha^{j_k,\ell,d=1} = \alpha^{j_k,\ell,d=0}$, $\ell \in \{1,\ldots,4\}$.[33] The only strong correlations are those between

[32] Appendix J presents the latent skill loadings on other types of tasks. Since we have 72 tasks in total, the tasks with the top 24 difficulty parameters are defined as easy tasks, the bottom 24 are defined as hard tasks, and the middle 24 are defined as medium tasks. All ranks are based on the estimates of the task difficulty level parameters.

[33] Tables I.3–I.4 in Appendix I provide item-by-item tests. Social-emotional item loadings are not

social-emotional skills and fine motor skills.

Table 9: Skill Loadings on Denver Test Tasks ($\alpha^{j_k,d}$) Latent Skills

Control			Treatment			p-value
Skill Loadings	Mean	S. D.	Skill Loadings	Mean	S.D.	for test of equality of means
Language and Cognitive	0.453	0.364	Language and Cognitive	0.679	0.469	0.000
Social-Emotional	0.259	0.263	Social-Emotional	0.222	0.246	0.002
Fine Motor	0.448	0.251	Fine Motor	0.556	0.211	0.001
Gross Motor	0.739	0.405	Gross Motor	0.693	0.442	0.276

Notes: 1. These are the means and standard deviations of $\alpha^{j_k,0}$ and $\alpha^{j_k,1}$, respectively, across items.
2. p-values are for the null of equality of treatment and control summary measures.

As is evident from equation (3), at the same level of skill, the larger the factor loadings, the better the child's performance on tests. Table 9 gives summary statistics (mean and standard deviations) for the skill loadings on different tasks. Except for gross motor skills, we reject equality of the summary statistics of treatment and control groups. In addition, the table shows the average effectiveness of each type of skill for performance on various tasks. For example, the loadings of latent language and cognitive skills are large for language and cognitive tasks, but the loadings of social-emotional skills for language and cognitive tasks are relatively small. This gives us some reassurance about the normalizations adopted.

3.2.3 Comparisons with a Model without Task Difficulty Parameters

To show the impact of introducing task difficulty parameters to the model, we estimate a restricted version of the model based on equation (3), in which we set all task difficulty parameters equal to zero. First, we compare the likelihood ratio between the full model and the restricted model and find that the full model has a higher likelihood. The likelihood ratio test statistic is $\chi^2(71) = 8419.26$, and the p-value of rejecting the null hypothesis of equal goodness of fit based on the two models is less than 0.001.

Second, we compare the treatment effects on the mean of latent skill factors in Table 10 ($E(\boldsymbol{\theta}^1) - E(\boldsymbol{\theta}^0)$). Notice that the estimates of a model without task difficulty parameters are very different from the estimates with the difficulty parameters. A model without difficulty parameters produces significantly negative

precisely estimated.

effects on social-emotional skills and significantly positive effects on gross motor skills, which are inconsistent with both the full model and the OLS model treatment effect evaluations.

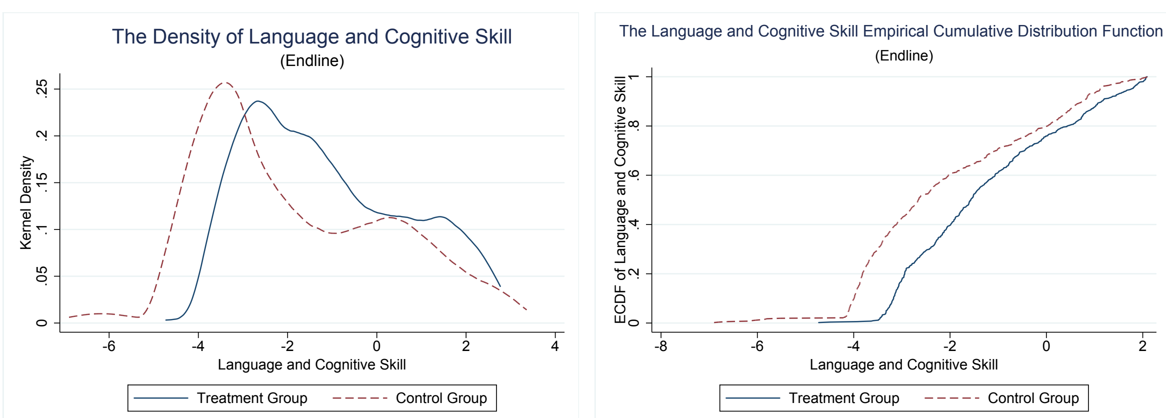
Table 10: Comparing Treatment Effects of $\boldsymbol{\theta_i}$ Based on Two Models with and without Difficulty Parameters

	Social-Emotional	Fine Motor	Language and Cognitive	Gross Motor
Full Model	0.395***	0.726***	0.753***	-0.095
(With Task Difficulty Adjustment)	[0.208, 0.583]	[0.551, 0.899]	[0.459, 1.051]	[-0.280, 0.089]
Restricted Model	-3.14***	1.136***	1.158***	1.069***
(Without Task Difficulty Adjustment)	[-3.375, -2.904]	[1.205, 1.505]	[0.857, 1.453]	[0.896, 1.237]

Notes: 1. The 95% confidence intervals in brackets are constructed by wild bootstrap clustered at the village level.
2. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.2.4 Distributions of Latent Skill

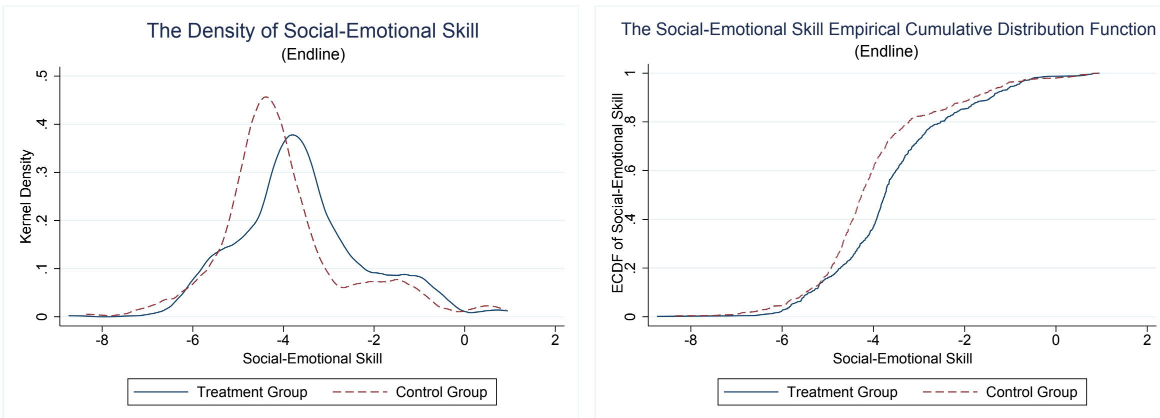
We compare the cognitive and language skill distributions of the control and treatment groups. Figure 5a shows that the density of language and cognitive skills for the treatment group shifts right and has a fatter upper tail than the one for the control group. Figure 5b shows that at almost every point of the cumulative distribution, language and cognitive skills are larger in the treated group than in the control group. Gains are more substantial for those who would be at the bottom and middle of the control distribution compared to those who would be at the top.



(a) (b)

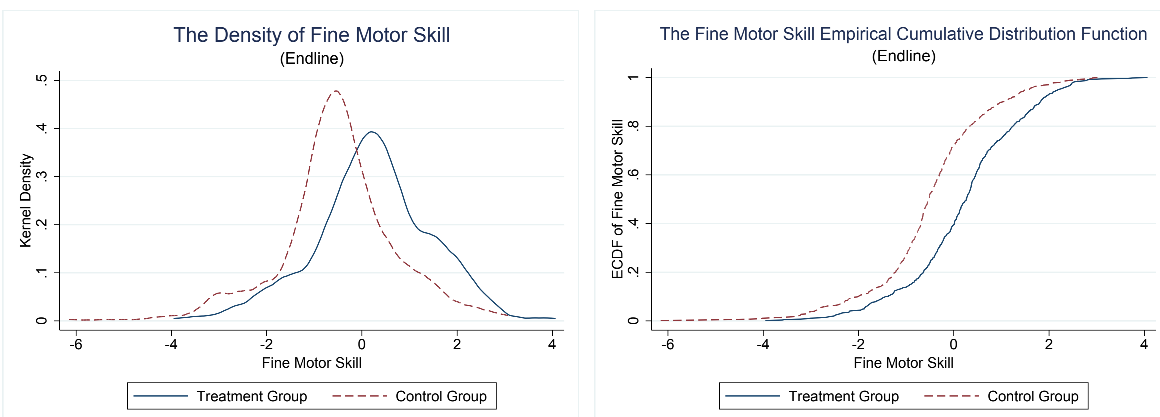
Figure 5: Language and Cognitive Skills Distribution

Figures 6a and 7a present the densities of social-emotional and fine motor skills, respectively. For social-emotional skills, gains are concentrated among those who would otherwise be at the center of the control distribution. For fine motor skills, gains are substantial throughout the entire control distribution.



(a) (b)

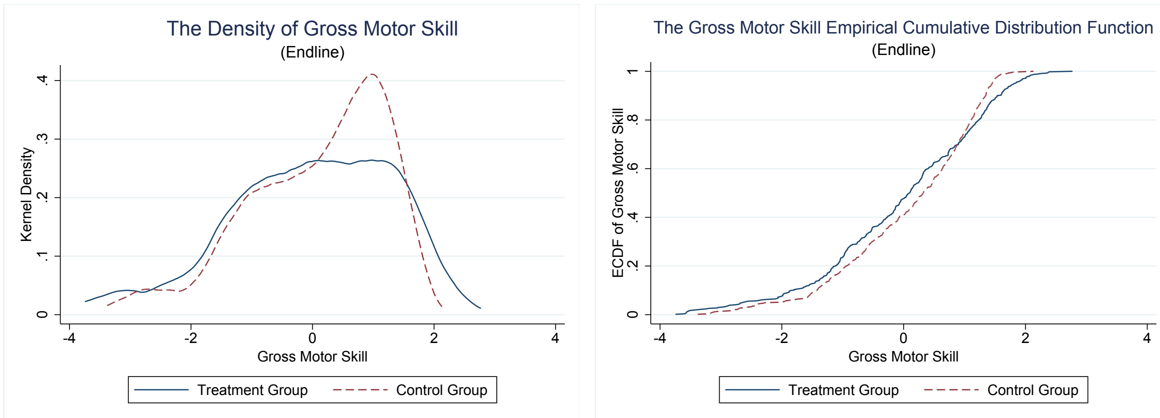
Figure 6: Social-Emotional Skills Distribution



(a) (b)

Figure 7: Fine Motor Skills Distribution

For gross motor skills, there is little evidence of any treatment effect. The factor distributions are similar between the control and treatment groups. Figures 8a and 8b show that the densities and CDFs of the two gross motor skills distributions are close to each other.



(a) (b)

Figure 8: Gross Motor Skills Distribution

In summary, language and cognitive, social-emotional, and fine motor skills were substantially improved by the program. But the gains are not uniform across the control distribution for cognitive skills. They are uniform for social-emotional and fine motor skills. Looking solely at mean treatment effects, we find significant improvements by the end of the intervention only in language and cognitive skills and not in fine motor and social-emotional skills. Examining the shift in the distribution of controls gives us a deeper look at who gains at which skill level. Appendix L presents an extension array of stochastic dominance tests for the estimated distributions.

4 Decomposing ATE

We use our estimates of latent skill profiles to understand the sources of the experimental ATEs. We compare experimental treatment effects with those obtained from our model.

4.1 The Sources of Treatment Effects

Average treatment effects produced by the experiment can arise either from changes in the mapping from skills to task performance or from changes in skills. We investigate the quantitative importance of each of these sources.

For each item j for skill k of the Denver test, the latent outcome for j is:

$$\begin{aligned}\tilde{Y}_i^{j_k} =& \boldsymbol{X_i'}\left[\boldsymbol{\beta}^{j_k,\mathbf{1}} D_i + \boldsymbol{\beta}^{j_k,\mathbf{0}}(1-D_i)\right] \\ &+ D_i(\boldsymbol{\theta_i^1})'\boldsymbol{\alpha}^{j_k,\mathbf{1}} + (1-D_i)(\boldsymbol{\theta_i^0})'\boldsymbol{\alpha}^{j_k,\mathbf{0}} + \varepsilon_i^{j_k}\end{aligned}$$

Since we recover the individual latent skills $\boldsymbol{\theta_i^d}$, we can use them as inputs into our estimates of equation (3) to simulate average treatment effects on Denver test scores. The point estimates of the average treatment effects so obtained are in close agreement.

Table 11: Average Treatment Effect Point Estimates Comparison

Denver Tasks	From OLS Model	From Factor Model	p-value
	ATE	ATE	
Language and Cognitive	1.113	1.115	0.504
	[0.723, 1.510]	[0.765, 1.454]	
Social-Emotional	-0.115	-0.081	0.556
	[-0.491, 0.275]	[-0.315, 0.152]	
Fine Motor	0.645	0.569	0.413
	[0.139, 1.158]	[0.136, 0.990]	
Gross Motor	0.219	0.190	0.460
	[-0.294, 0.775]	[-0.071, 0.450]	
$\chi^2(4) = 0.116$			0.998

Notes: 1. The 95% confidence intervals in brackets are constructed by wild bootstrap clustered at the village level.
2. The ATE estimates reported in this table are conditional on the pre-treatment covariates, which are consistent with column (5) of Table 2.
3. We conduct the Wald test to examine whether the two methods provide the same ATE estimates jointly. The p-value of the χ^2 test shows we cannot reject the hypothesis that the two methods produce the same ATE estimates.

4.2 Decomposing Treatment Effects

Experimental treatment effects may arise not only from enhancements of latent skills $\boldsymbol{\theta_i^d}$ but also from changes in the mapping from skills to task performance $\boldsymbol{\alpha}^{j_k,\boldsymbol{d}}$ and $\boldsymbol{\beta}^{j_k,\boldsymbol{d}}$. In order to understand the source of home visiting intervention treatment effects, we decompose the item-level treatment effects into two components: the effects from the changes in the mapping from skills to tasks and the effects of treatment on skills.

For each item j_k, the experimental outcome $Y_i^{j_k}$ is:

$$Y^{j_k}(d) = 1(\boldsymbol{X_i'}\boldsymbol{\beta^{j_k,d}} + \delta^{j_k} + (\boldsymbol{\theta_i^d})'\boldsymbol{\alpha^{j_k,d}} + \varepsilon_i^{j_k} \geq 0) \tag{5}$$

where we assume $\varepsilon_i^{j_k} \sim N(0,1)$. Home visiting treatment effects come from three channels: changes in the observable coefficient $\boldsymbol{\beta^{j_k,d}}$, changes in latent skill factors $(\boldsymbol{\theta_i^d})$, and changes in factor loadings for skills. Define $F^1(\boldsymbol{\theta^1}, \boldsymbol{X})$ and $F^0(\boldsymbol{\theta^0}, \boldsymbol{X})$ as the distributions of $(\boldsymbol{\theta^1}, \boldsymbol{X})$ and $(\boldsymbol{\theta^0}, \boldsymbol{X})$ in the treatment and control populations, respectively. Population treatment effects for item j_k can be decomposed as follows:

$$\begin{aligned}
&\Pr(Y^{j_k,1} = 1) - \Pr(Y^{j_k,0} = 1) \\
&= \underbrace{\int \{\Phi([\boldsymbol{X'}\boldsymbol{\beta^{j_k,1}} + \delta^{j_k} + (\boldsymbol{\theta^1})'\boldsymbol{\alpha^{j_k,1}}]) - \Phi([\boldsymbol{X'}\boldsymbol{\beta^{j_k,0}} + \delta^{j_k} + (\boldsymbol{\theta^1})'\boldsymbol{\alpha^{j_k,1}}])\} dF^1(\boldsymbol{\theta^1}, \boldsymbol{X})}_{\text{From Estimated Coefficients of } \boldsymbol{X}} \\
&+ \underbrace{\int \{\Phi([\boldsymbol{X'}\boldsymbol{\beta^{j_k,0}} + \delta^{j_k} + (\boldsymbol{\theta^1})'\boldsymbol{\alpha^{j_k,1}}]) - \Phi([\boldsymbol{X'}\boldsymbol{\beta^{j_k,0}} + \delta^{j_k} + (\boldsymbol{\theta^1})'\boldsymbol{\alpha^{j_k,0}}])\} dF^1(\boldsymbol{\theta^1}, \boldsymbol{X})}_{\text{From Latent Skill Loadings}} \\
&+ \underbrace{\int \Phi([\boldsymbol{X'}\boldsymbol{\beta^{j_k,0}} + \delta^{j_k} + (\boldsymbol{\theta^1})'\boldsymbol{\alpha^{j_k,0}}]) dF^1(\boldsymbol{\theta^1}, \boldsymbol{X}) - \int \Phi([\boldsymbol{X'}\boldsymbol{\beta^{j_k,0}} + \delta^{j_k} + (\boldsymbol{\theta^0})'\boldsymbol{\alpha^{j_k,0}}]) dF^0(\boldsymbol{\theta^0}, \boldsymbol{X})}_{\text{From Latent Skill Factors}}.
\end{aligned} \tag{6}$$

Notice that equation (6) holds over a common support for $\boldsymbol{X}$ and when the factors in the control and treatment groups have similar distributions of observable covariates, which is essentially satisfied in our sample.[34] Table 12 reports the decomposition of treatment effects. The main drivers of the treatment effects are increases in latent skills. We have shown that there is no significant difference on $\boldsymbol{\beta}$ between the treatment and control groups in Table 6. Therefore, the contribution on treatment effects from $\boldsymbol{\beta}$ is insignificant. The contributions from experimentally induced changes in $\boldsymbol{\alpha}$ are not precisely estimated. For this reason, we conclude that the dominant effect of treatment is on latent skills.

[34]To have a comparable sample between the control and treatment groups in our data, we restrict our sample to the children who are older than 12 months and younger than 46 months. In Appendix M, we show the age distribution between the treatment and control groups.

Table 12: Sources of the Treatment Effects

Tasks	Total Net Treatment Effects	From Observable Covariates	From Skill Loadings $\boldsymbol{\alpha}$	From Latent Skills $\boldsymbol{\theta}$
Language and Cognitive	1.096	-0.032	0.217	0.911
	(0.184)	(0.189)	(0.192)	(0.187)
		-3%	20%	83%
Social-Emotional	0.258	-0.001	0.049	0.211
	(0.082)	(0.086)	(0.088)	(0.084)
		-1%	19%	82%
Fine Motor	0.303	-0.009	-0.003	0.315
	(0.085)	(0.088)	(0.189)	(0.315)
		-3%	-1%	104%
Gross Motor	0.150	-0.028	0.062	0.117
	(0.098)	(0.105)	(0.109)	(0.102)
		-19%	41%	78%

Notes: 1. Total treatment effects for skill k are $\frac{1}{N_{J_k}} \sum_{j_k=1}^{N_{J_k}} \left(\frac{\sum_{i=1}^{N_I} Y^{j_k,i} D_i}{\sum_{i=1}^{N_I} D_i} - \frac{\sum_{i=1}^{N_I} Y^{j_k,i}(1-D_i)}{\sum_{i=1}^{N_I}(1-D_i)} \right)$ assuming both denominators are nonzero and N_I is # of observations.

2. To ensure that the observed covariates are balanced between the treatment and control groups, we consider the sample of children who are younger than 46 months and older than 12 months.

3. Standard errors are reported in parentheses.

4.3 Treatment Effects on Latent Skills Conditional on Caregiver Status

In this section, we compare the treatment effects based on the children's caregiver status. About 30–40% of children in our sample are left-behind children. Among the left-behind children, there are three cases: only father works outside, only mother works outside, and both parents work outside. Table 13 provides treatment effects on latent skill factors $\boldsymbol{\theta_i}$. Since the latent skill factors eliminate impacts due to task difficulty levels, the values are more comparable across different groups. Table 13 displays the strongest treatment effects for vulnerable children for whom mothers are absent (i.e., mother works outside or both parents work outside). Heckman and Zhou (2021) show that, in most cases, grandmothers with low levels of education are the caregivers when mothers are absent.

Table 13: Treatment Effects on Latent Skills $\boldsymbol{\theta_i}$

Standardized	(1) Non-Left-Behind Children	(2) Left-Behind Children: Mother Works Outside	(3) Left-Behind Children: Father Works Outside	(4) Left-Behind Children: Both Work Outside
Midline				
Language and Cognitive	0.503***	0.730**	0.308*	0.671*
	[0.258, 0.751]	[0.192, 1.330]	[-0.042, 0.661]	[0.049, 1.345]
Fine Motor	0.463***	0.555	0.669***	0.612
	[0.133, 0.797]	[-0.143, 1.246]	[0.225, 1.130]	[-0.143, 1.391]
Social-Emotional	0.453**	0.825	0.620**	0.622
	[0.075, 0.813]	[-0.174, 1.855]	[0.103, 1.156]	[-0.437, 1.596]
Gross Motor	-0.274**	-0.024	-0.292	-0.074
	[-0.494, -0.050]	[-0.581, 0.472]	[-0.692, 0.080]	[-0.681, 0.462]
Endline				
Language and Cognitive	0.539***	1.443***	0.828***	1.279**
	[0.125, 0.941]	[0.737, 2.255]	[0.456, 1.186]	[0.481, 2.150]
Fine Motor	0.619***	1.122***	0.831***	1.106***
	[0.428, 0.808]	[0.721, 1.499]	[0.477, 1.166]	[0.662, 1.519]
Social-Emotional	0.245*	0.311	0.560***	0.006
	[-0.013, 0.518]	[-0.283, 1.016]	[0.267, 0.867]	[-0.570, 0.649]
Gross Motor	0.114	-0.514	-0.320*	-0.448
	[-0.105, 0.339]	[-1.207, 0.104]	[-0.649, 0.008]	[-1.187, 0.247]
Pre-treatment Covariates	Yes	Yes	Yes	Yes
IPW	Yes	Yes	Yes	Yes

Notes: 1. The 95% confidence intervals in brackets are constructed using the wild bootstrap clustered at the village level.
2. The mean and variance for the standardized scores are estimated from the pooled sample of the control group children.
3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5 Comparison of China REACH Treatment Effects with Those of the Original Jamaica Reach Up and Learn Program

Table 14 shows that for comparable outcome measures at early ages, China REACH is on track with Jamaica Reach Up and Learn, which has been shown to generate substantial lifetime benefits (Grantham-McGregor and Smith, 2016; Gertler, Heckman, Pinto, Zanolini, Vermeersch, Walker, Chang, and Grantham-McGregor, 2014). We cannot reject the hypothesis that the treatment effects are the same across these two interventions. If China REACH continues on course, it should reproduce the effects of the successful Jamaica program.

Table 14: Treatment Effects on China REACH and Jamaica Reach Up and Learn

	Panel A: China REACH Latent Skill Factors (After 21 Months of Intervention)			
	Social-Emotional	Fine Motor	Language and Cognitive	Gross Motor
Treatment	0.40***	0.73***	0.75***	-0.10
	[0.21, 0.58]	[0.55, 0.90]	[0.46,1.05]	[-0.28, 0.09]
	Panel B: Jamaica Griffiths Test (After 24 Months of Intervention)			
	Performance	Fine Motor	Hearing and Speech	Gross Motor
Treatment	0.63***	0.67***	0.50***	0.34***
	[0.30, 0.95]	[0.34, 1.00]	[0.15,0.84]	[0.01, 0.67]
p-value	0.35	0.78	0.39	0.15

Notes: 1. For the China REACH program, the 95% confidence intervals in brackets are constructed by wild bootstrap clustered at the village level.

2. For the Jamaica Reach Up and Learn program, the 95% confidence intervals are presented in brackets.

3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4. The p-values in the last row correspond to the null of equality of treatment effects across the programs.

6 Conclusion

This paper estimates the impacts on child skills from a large-scale early childhood home visiting intervention program (China REACH). The program is patterned after the successful and widely-emulated Jamaica Reach Up and Learn program. Since national policy in China is driven by data, rigorous evidence on China REACH has the potential to have a large effect on policy discussions.

We estimate child latent skills and how they are affected by the program. We develop a framework for understanding the mechanisms generating treatment effects on child skill development that adjusts for the difficulty of the various tasks used to assess performance in the program. The program significantly improves child cognitive and language, fine motor, and social-emotional skills, but its impacts are not uniform across baseline skill levels. Its largest impacts are on the most vulnerable children. Improvements in latent skills explain the vast majority of estimated treatment effects. We test and reject the "dedicated factor" measurement model widely used in the economics of skill formation. Measured item scores depend on multiple skills. Our analysis offers a prototype for measuring latent skills using diverse outcome measures adjusting for the difficulty inherent in different tasks. Using these tools, we examine the impacts of skill interventions across baseline skill distributions.

References

ANDERSON, T. W., AND H. RUBIN (1956): "Statistical Inference in Factor Analysis," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, vol. 5, pp. 111–150, Berkeley, CA. University of California Press.

BAI, Y. (2019): "Optimality of Matched-Pair Designs in Randomized Controlled Trials," Unpublished manuscript, University of Chicago.

CAMERON, A. C., J. B. GELBACH, AND D. L. MILLER (2008): "Bootstrap-based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 90(3), 414–427.

CANAY, I. A., A. SANTOS, AND A. M. SHAIKH (2019): "The Wild Bootstrap with a "Small" Number of "Large" Clusters," *Review of Economics and Statistics*, pp. 1–45.

CHEN, M., I. FERNÁNDEZ-VAL, AND M. WEIDNER (2021): "Nonlinear Factor Models for Network and Panel Data," *Journal of Econometrics*, 220(2), 296–324.

CUNHA, F., J. J. HECKMAN, AND S. M. SCHENNACH (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78(3), 883–931.

ELANGO, S., J. L. GARCÍA, J. J. HECKMAN, AND A. HOJMAN (2016): "Early Childhood Education," in *Economics of Means-Tested Transfer Programs in the United States*, ed. by R. A. Moffitt, vol. 2, chap. 4, pp. 235–297. University of Chicago Press, Chicago.

GARCÍA, J. L., J. J. HECKMAN, AND A. L. ZIFF (2018): "Gender Differences in the Benefits of an Influential Early Childhood Program," *European Economics Review*, 109, 9–22.

GERTLER, P., J. J. HECKMAN, R. PINTO, A. ZANOLINI, C. VERMEERSCH, S. WALKER, S. CHANG, AND S. M. GRANTHAM-MCGREGOR (2014): "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica," *Science*, 344(6187), 998–1001.

GRANTHAM-MCGREGOR, S., AND J. A. SMITH (2016): "Extending The Jamaican Early Childhood Development Intervention," *Journal of Applied Research on Children: Informing Policy for Children at Risk*, 7(2).

HECKMAN, J. J., AND G. KARAPAKULA (2019): "Intergenerational and Intragenerational Externalities of the Perry Preschool Project," NBER Working Paper 25889.

HECKMAN, J. J., AND J. ZHOU (2021): "Interactions as Investments: The Microdynamics and Measurement of Early Childhood Learning," Unpublished Paper, University of Chicago.

HOMVEE (2020): "Early Childhood Home Visiting Models: Reviewing Evidence of Effectiveness, 2011-2020," OPRE Report 2020-126.

HOWARD, K. S., AND J. BROOKS-GUNN (2009): "The Role of Home-Visiting Programs in Preventing Child Abuse and Neglect," *The Future of Children*, 19(2), 119–146.

LIZZERI, A., AND M. SINISCALCHI (2008): "Parental Guidance and Supervised Learning," *Quarterly Journal of Economics*, 123(3), 1161–1195.

LU, B., R. GREEVY, X. XU, AND C. BECK (2011): "Optimal Nonbipartite Matching and Its Statistical Applications," *American Statistics*, 65(1), 21–30.

MAASOUMI, E., AND L. WANG (2019): "The Gender Gap between Earnings Distributions," *Journal of Political Economy*, 127(5), 2438–2504.

RYU, S. H., AND Y.-J. SIM (2019): "The Validity and Reliability of DDST II and Bayley III in Children with Language Development Delay," *Neurology Asia*, 24(4), 355–361.

TSIATIS, A. (2006): *Semiparametric Theory and Missing Data*. New York: Springer.

VAN DER LINDEN, W. J. (2016): *Handbook of Item Response Theory: Volume 1: Models*. CRC Press.

WANG, F. (2020): "Maximum likelihood estimation and inference for high dimensional generalized factor models with application to factor-augmented regressions," *Journal of Econometrics*.

Online Appendix

The Impacts of a Prototypical Home Visiting Program on Child Skills

Jin Zhou, James Heckman, Bei Liu, and Mai Lu

February 9, 2022

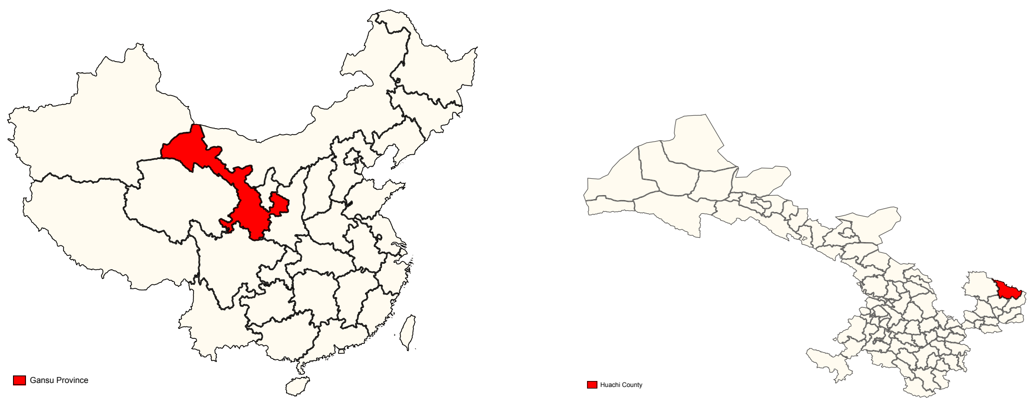
Contents

List of Figures

List of Tables

A Conducting the Experiment

Field efforts included recruiting and training interviewers, hiring county-based Ph.D level project directors, and engaging support by the local government and health system. In early 2015, local home visitors were recruited and trained. At the same time, in January 2015, the baseline data were collected, which covered all of the presented households with children under 2 years of age in Huachi county. The information includes the Infant-Toddler HOME Inventory for 0-36 months, a household demographic survey, and village-level registration data. Fifty-six villages are selected into the treatment group, and fifty-five villages are in the control group. Figure A.2 shows the locations of treatment and control group villages in Huachi County.



(a) Gansu Province (b) Huachi County

Figure A.1: The Location of Huachi County

In January 2016, there was a sample enlargement, including children born between February 9th, 2015, and April 30th, 2015, for both control and treatment groups. Newly-enrolled children were assessed on their health (height, weight, head circumstances, hemoglobin), the Denver II test, the Infant-Toddler HOME Inventory for 0-36 months, and a household demographic survey before the home visiting intervention began.

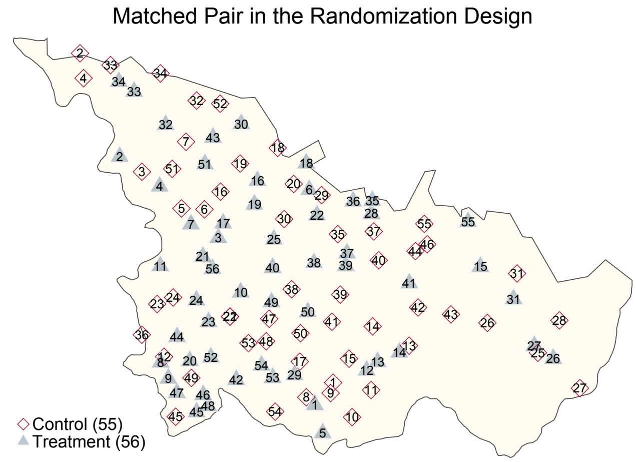


Figure A.2: The Locations of Randomized Paired Villages in Huachi County

From September 2015 to November 2017, home visitors provided weekly one-hour home visits to the treated children. Home visitors taught caregivers to encourage infants' and toddlers' play-based learning during daily family activities and routine events. Culturally adapted learning materials such as picture books, games, and toys were employed to demonstrate specific interactions.

Between September 2015 and November 2017, two rounds of follow-ups were conducted in July 2016 and July 2017, respectively, which include early childhood development assessments: anthropometrics (height, weight, head circumference, and hemoglobin), the Denver Development Screening Test (2nd ed) for 6-72 months, the Infant-Toddler HOME Inventory for 0-36 months, Early Childhood HOME (36-60 months), and a household demographic survey.

January 2015 Before home visit interventions, baseline information were collected for children with birth dates between April 1, 2013, and November 30, 2014.

September 2015 Initiation of the home visit.

January 2016 Enlarging the enrollment:[1]

- Including 180 children (born between February 9, 2015, and April 30, 2015) for both treatment and control groups.
- Assessed children's health and Denver II scores before home visit intervention.

[1] By January 2016, the data sample included 1566 children: 247 attrited and 76 were newly enrolled. In Section G, we document how we deal with the missing data problem.

July 2016 Conducted a family survey and assessed children's health and Denver II test scores for both control and treatment groups.

July 2017 Conducted a family survey and assessed children's health and Denver II scores for both control and treatment groups.

November 2017 End of home visit intervention.

B China REACH Program Home Visitor Guideline

B.1 About China REACH

The following appendix is an English translation of the guideline book for home visitors in the China REACH program are trained with. It documents how home visitors should approach their work with families and ways to encourage healthy relationships between caregivers and children. It also touches on advice home visitors can give to caregivers in order to strengthen these attachments.

B.1.1 Intervention Goals

In the course training, home visitors learn the process of child development and how caregivers can promote child development. The home visitors also learn how to conduct home visits and show caregivers how to initiate warm and supportive interactions with their children. Using role-playing and toys, they practice how to work with caregivers to facilitate these interactions.

Curriculum goals for the home visitor:

- Understand the role played by the home visitor program and the home visitor.
- Learn about child development.
- Learn about the skills of conducting home visits and how to introduce new skills and activities to families.
- Learn how to record home visits throughout the course progress.
- Make toys for home visits.

Learning goals for caregivers in the intervention:

- Acquire knowledge about child development.
- Improve ways of talking, playing games, teaching, and interacting with children.
- Be able to use daily activities and household items to teach and play with children.
- Improve self-confidence and gain happiness in the process of promoting children's growth.

The learning goals for children in the project:

- Improve language ability and advance intellectual development.
- Improve behaviors and develop their social and emotional abilities.

B.1.2 Children's Intellectual Development

With the help of adults, children explore things in their environment, play with and communicate with adults, and imitate their behaviors.

- Children aged 6-24 months

  Adults show children how to interact with the following materials:
  - Items of different colors, shapes, and materials.
  - Items that can be put in or taken out, items that make a sound when colliding, items that can be stacked together, and items that can be opened and closed.
- Children aged 24-48 months

  Children learn from the following activities:
  - Playing games
  - Imitating, pretending to be an adult
  - Playing puzzle games
  - Building objects and matching them. Playing games is very important for children, and it is essential for their intellectual development. When adults are involved and describe what the children are doing, they learn the most and are the most effective.

B.1.3 How to Improve Children's Language Skills

The home visitor aims to help the caregiver support the child's language development by facilitating interactions that focus on the following methods:

- Respond to the child's voice, words and questions.
- Introduce new sounds and words.
- Talk to your children as much as possible and describe what you are doing (for example: "I am making breakfast for you now," "I am washing these dirty clothes.")
- Call out the names of objects and people at home and outside.
- Tell the children what they are holding, what they want to reach, and what they are interested in.
- Do a naming game: say the name of the object and let the child point it. Then point to the item and ask the child to call out the item's name.
- Give praise to your child when he/she uses new words correctly.
- Expand the child's vocabulary (e.g., When the child says, "Look, dog," the caregiver says, "Yes, this is a big white dog.")
- Watch albums and photos with them.
- Use situations in daily life (such as bathing, coaxing him/her to sleep, eating, etc.) to speak new words and words learned before.
- Play role-playing games with your child (for example, say "Do you like your baby? Let's feed her something.")

B.1.4 How to Help Children Develop Their Social and Emotional Abilities

The child needs a caregiver that is:

- Someone they trust (they respond to their needs, such as hungry, unhappiness, etc.).
- Someone who expresses love to them (such as hugging, comforting, kissing them, speaking to them softly and gently, telling them that they love them).
- Someone who communicates with them (responds to their voice).
- Someone who understands them (knows what they like to do, what makes them happy/unhappy), and plays with them.
- Someone who is always by their side (the caregiver they have always been familiar with).

This will make them feel safe, confident, and happy. They will develop a strong, secure attachment to this person, and the strength of this attachment will affect the way they get along with other people in the future and their happiness in later life.

B.1.5 How to Improve Children's Self-Confidence

Children living in poverty usually lack self-confidence. The following activities can help them strengthen their self-confidence.

- Refer to the children using their names as much as possible.
- Praise the children on their personal images/behaviors (For example, "You are already a big kid," "You are so smart," "You are a helpful kid").
- Discuss with them on what they are doing ("I see you flipping through the picture album and eating," "I see you like this doll/that toy").
- Always listen and respond to the child's voice.
- Allow him/her to make their own choices.
- Give them praise when they successfully complete tasks or play games well.
- Make sure that the children experience successes more often instead of failures.
- Ask the caregiver to make a toy bag for the child, put it somewhere in the house, and then tell the kid that the bag belongs to him/her.

What will hurt children's self-confidence:

- Regular punishments, especially beating, ridiculing, and blaming.
- Frequent failure experience and/or often giving the children difficult tasks.

B.1.6 Home Visit Guide

The home visitor helps the caregiver introduce new activities to the child during visits. Activities such as making toys, completing puzzles, and making picture albums are part of what the Home visitor brings to the visit with each family. The goal is to have the caregiver interact warmly, and spontaneously with the child. A home visitor is assigned to each family at enrollment and she continues to see the family for the duration of the program. This helps build a relationship with the family, which in turn helps form trust.

- It is very important not to change the home visitor of each family in the project.
- It is also very important to complete each home visit. The higher the frequency of home visits, the better the child's growth.

B.1.7 Caregiver's Role

Caregivers play a crucial role in their child's developmental success. When a caregiver learns how to support his or her growing child, that knowledge can be beneficial long after the intervention is over.

Home visitors lead home visits, but caregivers and home visitors share responsibilities. Caregivers eventually take over the duties of home visitors.

We need to encourage caregivers to decorate the home into a more stimulating environment, play with children with home and outside objects, talk to and play with children in daily activities, teach them new knowledge, and spend time with them to play with toys, look at the picture album and communicate with words.

B.2 Training Sessions

B.2.1 Home Visit

- How does the home visitor establish this positive relationship?
  - Sits upright
  - Asks the caregiver what activities and games she has done with her child recently
  - Actively listens (recognizes and thinks about what the caregiver said)
  - Praises her
- How do you tell if the caregiver is satisfied with this home visit?
  If the caregiver
  - Smiles or laughs
  - Is able to complete the activity well
  - Plays with kids with pleasure
- What does the home visitor do to make the caregiver satisfied with the home visit?
  - Praises the caregiver
  - Praises the child
  - Makes sure the caregiver and the child know they are doing well

B.2.2 How to Help Children Learn

- Background information
  - Children learn knowledge and skills through imitating, exploring and experimenting.
  - Children learn through multiple repetitions.
  - Children learn the best when adults are involved.
  - When introducing a new activity, find out the most accessible part for the child.
- How to help children learn
  - Give the child new materials and allow him enough time to explore the materials independently. Observe what he did and praise it (for example, "Wow, you picked up a building block. Look, he picked up a building block, so smart.").
  - Explain the goal of the new activity to the caregiver (e.g., put the blocks into and out of the container).
  - Demonstrate new activities to children.
  - Encourage children to do activities and allow them to practice independently. Praise her when she tries to perform activities.
  - Involve mothers in children's activities.
  - Whenever the caregiver tries to participate in activities, remember to praise her.
  - Let children practice more independently.
  - When mothers and children participate in activities, praise them.

B.2.3 How to Teach Mothers

- How do you tell if the caregiver has understood how to do this activity with the child? The caregiver:
  - Is able to do this activity with children
  - Praises the child
- How does the caregiver know what she should do? The home visitor:
-
  - Explains clearly
  - Demonstrates activities
- How do you judge mothers and children like this activity? The caregiver and child:
  - Are happy
  - Can perform this activity well
- What makes them like this activity?

- The home visitor praises caregivers and children, demonstrates activities, and gives caregivers time to practice.

B.2.4 The Importance of Praise/How to Make Home Visits Interesting

- Praise can help the parenting home visitor to establish a positive relationship with the caregivers and help them build self-confidence.
- Praise can also improve children's language skills, promote the development of social and emotional skills and self-confidence.
- Praising a child can make him feel smart and willing to do this activity.
- Compliment your child in the following situations:
  - When playing with toys
  - When communicating with movement, voice, or speech
  - When having fun during a home visit
  - When trying an activity, even when it is unsuccessful
  - When performing activities
- Praise can be verbal or non-verbal. Non-verbal praise includes applause, high five, hug, smile etc.
- Importance of increasing the fun of home visits
  - Better establish a relationship with caregiver and child
  - Enhance the confidence of caregivers and children
  - Make caregivers feel more comfortable and natural
  - Let children feel at ease, confident and happy
  - Motivate children to learn
  - Increase children's participation

B.2.5 Listen, Understand, and Respond to Children

- Young children communicate with others by pointing, making noise, reaching out, crying, and smiling. As children grow older, they begin to speak a few words and then a few sentences.
- Explain to the caregiver that they should always tell their children what they are doing and what they want to express.

B.2.6 Build a Good Relationship with Caregiver

- The importance of establishing a good relationship with the caregiver. Good relationships

  - Encourage her to participate in activities
  - Build caregivers' confidence
  - Make caregivers more willing to listen to the advice of a parenting counselor
  - Make caregivers more willing to share personal difficulties and successes
- Home visitors and caregivers establish a good relationship through the following methods:
  - Sit upright
  - Ask about activities that have been done with the child
  - Actively listen (recognize and think about what mom said)
  - Praise her

B.2.7 Understand Difficulty

- It is very important to give the child activities that match his ability. If it is too easy, the child will feel bored. If it is too difficult, the child will feel frustrated and shocked.
- If the task is difficult to complete, you can break it down into simple steps or reduce the difficulty.
- If your child can complete simpler tasks, try to increase the difficulty.
- It is important to ensure that children understand and are competent in simple activities when they start higher-level games.
- When the child himself can repeat an activity correctly without the help of others, we can make it more difficult for him.

B.2.8 Give Feedback to the Caregiver

- When caregivers are involved in a certain activity or perform well, the home visitor should explain what they are doing well.
- Feedback should also make caregivers feel more comfortable during home visits and give them enough confidence to participate in activities on their own.
- Praise the caregiver

  Always praise the caregivers. You can praise them in the following situations: when they
  - Participate in activities with their children
  - Praise their children
  - Chat with their children
  - Tell you what their children did/can do
  - Tell you what activities they and their children did (example)

- You can praise the caregiver in these ways:
  - Recognize what the caregiver has done
  - Smile at the caregiver
  - Use specific praise to let the caregivers know what they are doing well

B.2.9 Use Daily Activities to Help Children Learn

Children learn by observing and imitating others. They will eventually want to do things on their own and explore their surroundings. Caregivers can teach them new things through daily activities, talking, and playing with the children. Caregivers should be encouraged to use things at home and outside for children to play.

B.2.10 First Home Visit

The first home visit is especially important because parents must understand the project's goals, the content of the activities, and what they are going to do.

- The home visitor starts the home visit by greeting the caregiver and introducing herself. They should both understand how to address each other.
- The home visitor asks the caregiver about the child:
  - The home visitor explains the project
  - The home visitor asks about family information and who lives in the home. She also notes whether there are other children in the family and their ages.
  - Before the end of the visit, the home visitor should make an appointment with the caregiver for the next visit
- Explain the project
  - Number of home visits
  - What you will do and its importance
  - Impact on child development
- Ask about family information and who lives at home. Note whether there are other children in the family and their age.
- Make an appointment with your caregiver for the next home visit.
  - Arrange a suitable time for you, the caregiver, and the child.
  - Record the phone number or other contact information.

B.2.11 Involving Other Family Members

Family members other than the primary caregiver in the program are very important in a child's life.

- Some fathers and grandparents showed great interest in the intervention, and their support was very helpful. Including these family members will contribute positively to the child's experience.
- Include any adults living in the family, such as fathers or grandparents, who are willing to take care of their children and participate in activities.
- Bring toys (picture albums/crayons and paintbrushes) to other children living in the home for their play. Some activities, such as games, are more appropriate for other children to include.
- Why is it important for other family members to include?

  When including family members, children will:
  - Feel happy
  - Have more fun
  - Build confidence

B.2.12 Promote Positive Behavior

Children learn by observing and imitating others. They will eventually want to do things themselves and explore their environment. Good management of children's behavior will make the activity smoother and make everyone feel more comfortable.

Here are several ways to manage your child's behavior. We can play with children, praise them, give them choices, keep them safe and distract them.

We can praise children in these situations: when they

- Do something well
- Try to do something (even if they do it wrong)
- Show the real themselves

Giving children the right to choose makes them feel that they can control what happens. For example, allow children to choose which book to read or which food to eat.

As children grow up and become more independent, they will want to explore more. Avoid always saying "no" to your child. One way is to make the home environment safer for children to explore. Put away what the child can reach and break.

Distract children and keep them away from things they cannot touch or things they cannot do instead of saying "no."

B.3 Curriculum

The development of skills in young children has been extensively studied and theorized over the years (e.g., Uzgiris and Hunt (1975) and Palmer (1971) are major references). The

China REACH program curriculum is adapted from the Jamaican Reach Up and Learn program, which is designed to focus on a child's ability to complete sequences of tasks ordered by progression difficulty levels based on general child development patterns. In general, children's skill development depends on a number of factors such as caregiver involvement, cultural environment, nutrition, child endowment, etc. To better understand how the skills develop over time, it is necessary to analyze the measures used to evaluate children's multidimensional skills. Based on the main content of tasks, the tasks in the curriculum cover four domains of skills.[2] The categories help researchers understand how the main type of skills developed based on the measures in the curriculum. Next, we document all the tasks in the China REACH curriculum by four domains of skill types. Next, we document all the tasks in the China REACH curriculum by four domains of skill types: fine motor, gross motor, language, and cognitive skills.

B.3.1 Skills Taught in the Curriculum

Fine motor, gross motor, language, and cognitive skills are taught. Within each skill group, skills are ordered by difficulty level following the patterns developed by Palmer (1971). Skills are sorted into different difficulty levels. For example, there are seven difficulty levels for fine motor drawing lessons.for[3,4] In general, the higher difficulty level for skills includes new content. For example, difficulty level 2 is to mimic circles. The skills at difficulty level 3 include drawing straight lines. We document how the tasks into different difficulty levels are categorized.

For example, Fine Motor Drawing lessons focus on a child's ability to use writing utensils with increasing skills. First, a child is asked to hold the utensil to make markings. Next, the child should incorporate more and more cognitive skills to complete the tasks. They then begin by copying markings made by an adult. As skill levels progress, they are asked to make the marking after only a verbal command from the adults. Finally, the child progresses from abstract shapes to representative drawings (See Table B.1.).

[2]We also are aware that skills do not develop in isolation, fine motor skills require cognitive input and language skills develop in tandem with gross motor functions.

[3]The standard of generating the difficulty levels are based on the understanding of the content in the skills.

[4]The difficulty level in our content only has ordinal meaning, not cardinal meaning.

Table B.1: Skill Levels for Fine Motor (Drawing) Lessons

Difficulty Level	**Task Content**
Level 1	Doodle using crayons
Level 2	Mimic draw circles
Level 3	Mimic circles and draw straight lines
Level 4	Draw a circle, vertical line, and horizontal line
Level 5	Draw circles, many lines, and crossed lines
Level 6	Draw a cross (or T), curves, and zigzag curves
Level 7	Draw caterpillars

In addition to tasks of different difficulty levels, the curriculum features multiple lessons and assessments at the same difficulty level l. The difficulty level category descriptions are listed in this section. The number of lessons within each difficulty level depends on the content in the curriculum. For example, there are six assessments at difficulty level 3 for fine motor drawing skills and only two assessments at difficulty level 2.

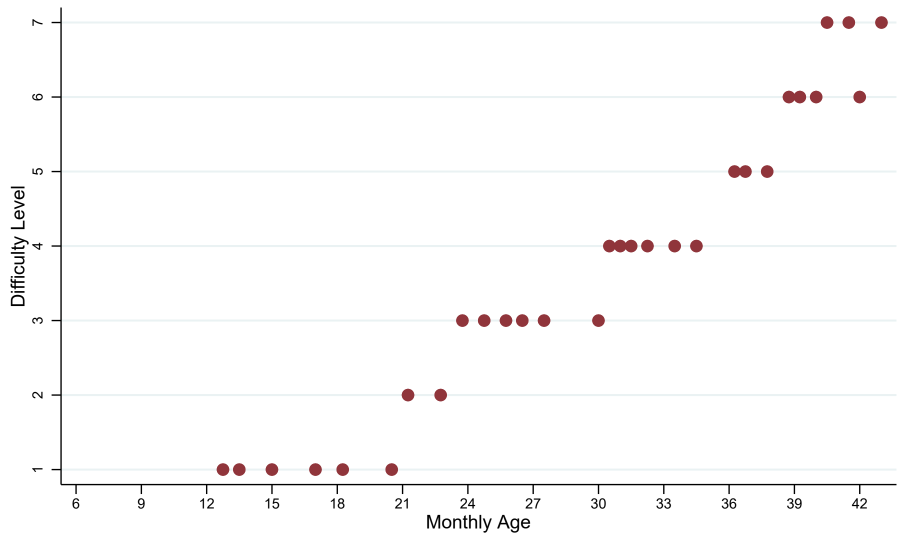
Figure B.1 shows the timing of each fine motor drawing assessment in the curriculum design. For the designated skills, difficulty level 1 covers from 12 months and 3 weeks to 20 months and two weeks. This timing means that when the child is 12 months and three weeks old, the home visitor will teach her the first fine motor drawing skill. When she is 20 months and two weeks old, the home visitor will teach her the sixth lesson at difficulty level 1. In general, higher difficulty levels appear at later weekly ages. However, there can be some overlap across difficulty levels. For example, in Figure 2, by the time difficulty level 7 of fine motor lessons start, the last lesson of level 6 remains unfinished. In Figure B.1, when fine motor lessons at difficulty level 7 start, the student still receives lessons at difficulty level 6. Circling back is a strategy designed to solidify a child's understanding of a concept.

Another example concerns cognitive skill categories. Cognitive skills have different dimensions. In the curriculum, the cognitive skills taught cover spatial, knowledge of objects and object functions, order and number, etc. Using knowing objects and object functions as an example: cognitive skill difficulty levels are defined based on the abstract concepts shown in Table B.2, such as the child's proficiency in understanding the objects. Seventy-four lessons are sorted into the listed 13 ordered difficulty levels.[5] It covers the process of how the child learns to know an object and understand the function of the object.

The lessons in the cognitive knowledge of objects unit progress from a simple understanding of the concept of pictures by acknowledging with vocalizations, to using receptive (heard) language to identify certain pictures. Receptive language is a skill developed prior to

[5]The difficulty level in our content only has ordinal meaning, not cardinal meaning.

Figure B.1: The Timing of Fine Motor Skill (Drawing) Tasks across Difficulty Levels



Note: Level 1: Doodle using crayons; Level 7: Draw caterpillars.

the expressive language where a child forms words to communicate. The child must use his or her expressive language to complete the following lessons, which increase with difficulty as they must develop more and more language to identify an increasing number of images. To progress through level 7 and beyond, the child must display an increasingly sophisticated understanding of the stories presented, first simply naming actions, then answering questions, then talking abstractly about a story. Levels 10, 11, 12, and 13 ask the child to take the information presented and build on it by discussing the uses of objects presented and making connections with other images.

Figure B.2 shows the timing of each cognitive (knowing objects and understanding the object's function) level in the curriculum. According to the curriculum content, the number of lessons varies across difficulty levels. Table B.3 presents detailed information about the six lessons (and assessments) that are labeled as difficulty level one directed to ten-month to 15 month-old curriculum content. In Table B.3, all lessons relate to the activity of looking at the pictures or objects and vocalizing, which does not require the child to name or identify the object.

Table B.2: Difficulty Level List for the Cognitive Understanding Objects Lessons

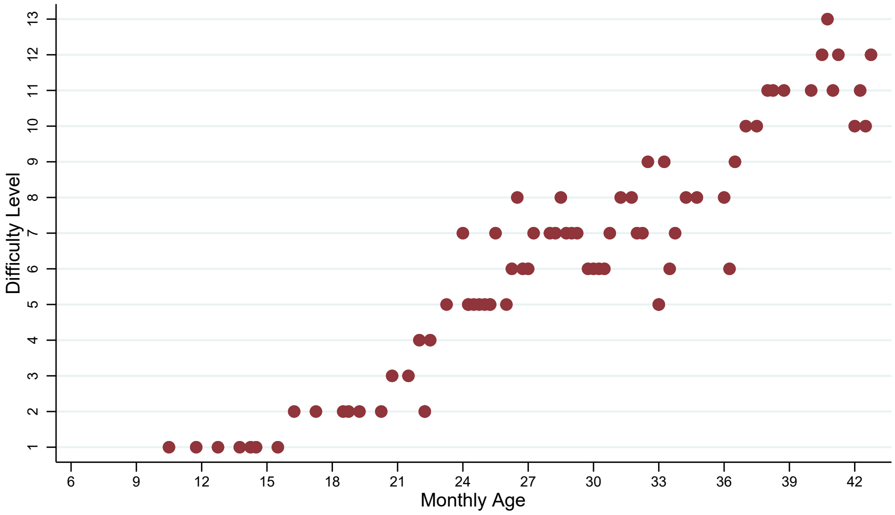
Level 1	The child can look at the pictures and vocalize
Level 2	Name the objects and ask the child to point to the corresponding pictures
Level 3	The child can name the objects in one picture, and point to the named picture
Level 4	The child can name the objects in two or more pictures, and point to the named picture
Level 5	The child can point out named pictures, and say names of three or more
Level 6	The child can point out the picture mentioned and correctly name the name of 6 or more pictures
Level 7	The child can talk about the pictures, answer questions, understand, or name the verbs (eat, play, etc.)
Level 8	The child can follow the storyline, name actions, and answer question
Level 9	The child can understand stories and talk about the content in the pictures
Level 10	The child can keep up with the development of the story
Level 11	The child can say the name of each graphic, discuss the role of each item, and then link the graphics in the card together
Level 12	The child can name the things in the picture, link different pictures together, and discuss some of the activities in the pictures
Level 13	The child can name the things in the picture and talk about the function of objects

Table B.3: Cognitive Skill Task Content: Look at the Pictures and Vocalize (Level 1)

Difficulty Level	Month	Week	Learning Materials	Content
1	10	2	Picture book A	The baby makes sounds when looking at the pictures
1	11	3	Picture book B	The baby looks at the pictures and vocalizes
1	12	3	Picture book A	The child makes sounds looking at the pictures
1	13	3	Picture book B	The child makes sounds looking at the pictures
1	14	1	Picture book A	Mother and child look at the pictures together, and the mother lets the child vocalize and touch the pictures
1	15	2	Picture book B	Mother and child look at the pictures together, and the mother lets the child vocalize and touch the pictures

In sum, the curriculum targets lessons at different skill levels for multiple levels of skill at each weekly age. For each type of skill, the task difficulty levels based on the content of the tasks and the guideline of Uzgiris and Hunt (1975) and Palmer (1971) are constructed. The terms of the number of lessons within each difficulty level varies. We follow these scholars and assume that each level is a quantum of understanding comparable across children. We use achievement at each level of skill as our measure of knowledge.

Figure B.2: The Timing of Cognitive Skill (Understanding Objects) Tasks across Difficulty Levels



Note: Level 1: Look at the pictures and vocalize; Level 13: The child can name the things in the picture and talk about the function of objects.

B.3.2 Fine Motor Skill

The fine motor skill involves finger movements, such as grasping, releasing and stitching, and drawing and writing skills. Here we consider two types of fine motor skills: (1) finger movements related to grasping, releasing, stitching; and (2) the movements related to drawing and writing ability. This task evaluates whether a child can grasp the writing instrument and make marks, scribbles, and shapes. It is not writing ability as in letters or words.

The first category is related to finger movements regarding grasping, releasing, stitching.[6] In Table B.4, tasks progress from basic activities like holding and moving an object that require limited precision with the fine muscles of the hands to manipulating the object with movements that need incrementally more dexterity (like rotating the object) to complex tasks requiring finer and finer finger control, like unscrewing the top. Finally, tasks that require the most hand dexterity, as well as hand-eye coordination, come last.

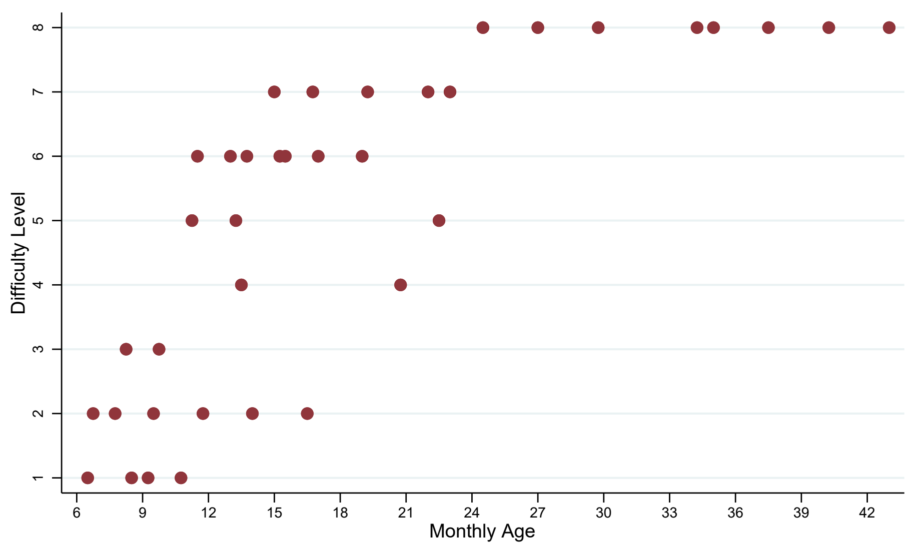
[6] These milestones are justified at `https://www.chrichmond.org/therapy-services/occupational-therapy/developmental-milestones/fine-motor-skills-birth-to-2-years` and `http://www.kamloopschildrenstherapy.org/fine-motor-skills-infant-milestons`.

Table B.4: Difficulty Level List for Finger Movement Tasks

Level 1	Rattle the bottle
Level 2	Shake and beat the drum with two hands
Level 3	Pull strings to get toy
Level 4	Rotate, push
Level 5	Place small objects into the bottle, shake it, and unscrew the lid
Level 6	Put small container into a larger container
Level 7	Take the ring off and slip the ring onto the bottle
Level 8	String beads

Figure B.3 gives the timing of each finger movement tasks in the curriculum.

Figure B.3: The Timing of Fine Motor Skill (Grasping, Releasing Actions) Tasks across Difficulty Levels



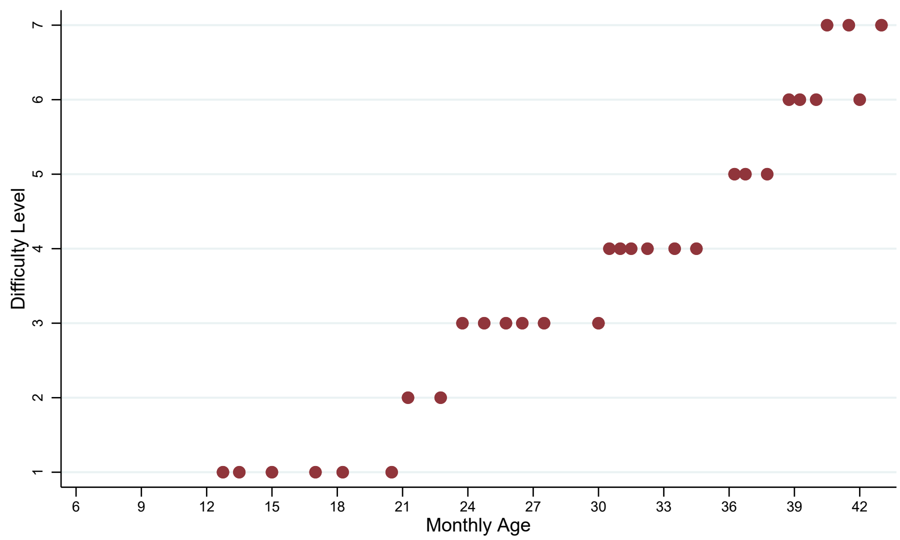
Note: Level 1: Rattle; Level 8: String beads.

The second category is related to drawing and manual writing ability. The fine motor drawing tasks in Table B.5 focus on a child's ability to use a writing tool with increasing skills. First, a child must be able to hold the tool to make markings. Next, the child must incorporate increasingly complex cognitive skills to complete the tasks. They start by imitating markings made by an adult. Then, when skill levels progress, they must make the marking after only a verbal command from the adult. Finally, the child progresses from abstract shapes to representative drawings.

Table B.5: Difficulty Level List for Fine Motor Drawing Tasks

Level 1	Doodle using crayons
Level 2	Mimic draw circles
Level 3	Mimic circles and draw straight lines
Level 4	Draw a circle, vertical line, and horizontal line
Level 5	Draw circles, many lines, and crossed lines
Level 6	Draw a cross (or T), curves, and zigzag curves
Level 7	Draw caterpillars

Figure B.4: The Timing of Fine Motor Skill (Drawing) Tasks across Difficulty Levels



Note: Level 1: Doodle using crayons; Level 7: Draw caterpillars.

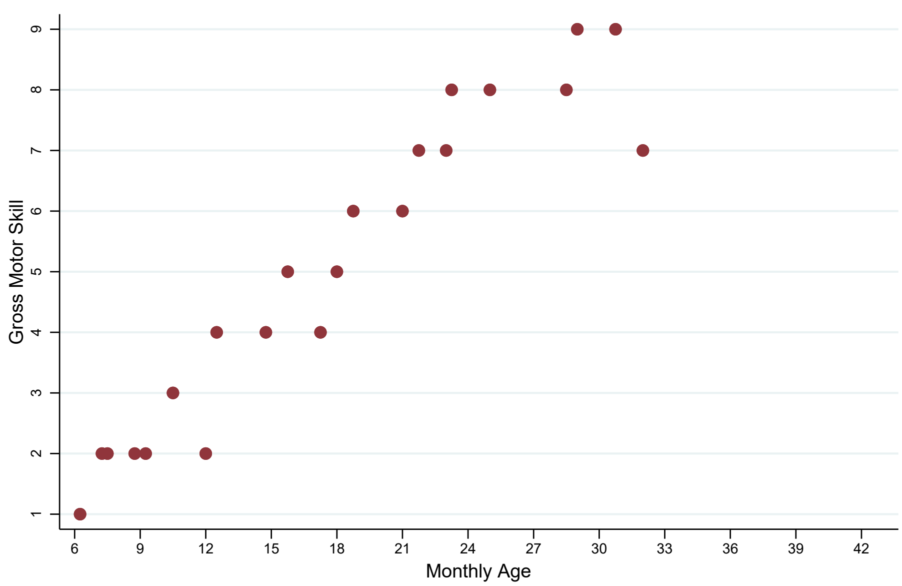
B.3.3 Gross Motor Skill

Gross motor skill is any skill that requires movement and precision of large muscles in the body. Crawling, creeping, walking, throwing and dancing are all examples of gross motor skills. The designated gross motor tasks start with a relatively simple activity, touching the ball, requiring the child only to move one hand to the object. Next, the child must be able to move his or her entire body to interact with the toy. After mastery over those tasks, the child uses both gross motor skills and newly found cognitive ability to interact with the toy in increasingly complex ways. Pushing a toy requires coordination, standing, and walking skills. However, the child is still using the toy as a walking aid at this point. To progress to the next tasks, not only will the child have to master walking independently, but will also use the toy in a way that suggests intentionality (e.g., pulling, throwing). The final tasks require the child to integrate cognitive knowledge of direction, descriptive words, and gross motor mastery of balance.

Table B.6: Difficulty Level List for Gross Motor Tasks

Level 1	Let the child touch the ball
Level 2	The child moves (crawls) and follows the ball
Level 3	Roll the ball
Level 4	Push the toy when walking
Level 5	Pull the toy
Level 6	Pull and walk forward or backward
Level 7	Throw ball backward, forward, upward and into a target
Level 8	Move forward or backward. Child can understand "upward," "downward," "inside of," "outside of," "stop," "go," "fast," "slow."
Level 9	Hold the soft ball on his or her head stably while walking

Figure B.5: The Timing of Gross Motor Skill Tasks across Difficulty Levels



Note: Level 1: Let the child touch the ball, Level 9: Hold the soft cloth ball on his head stably while walking.

B.3.4 Cognitive Skill

Cognitive skill is broadly defined as a child's ability to apply what they have learned previously for new situations. This skill involves logic, problem-solving ability, memory, attention, and so on.

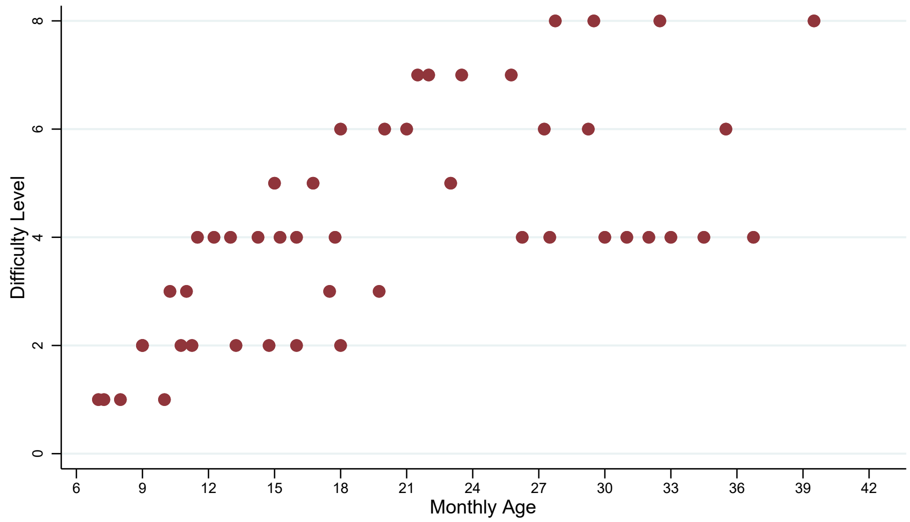
B.3.4.1 Spatial Skills

Spatial skills rely on a child's understanding of the three dimensional world. Comprehending concepts of relative positioning—"inside of," "around," and "next to" are the basics of this skill. The progression of these skills follows the child as he or she learns concepts that are more and more abstract. Beginning with "in" and "out" and progressing to "underneath," "around," "up," "next to," and "close to." As the tasks become more difficult, the child is expected to manipulate objects to demonstrate knowledge and understanding of these concepts.

Table B.7: Difficulty Level List for Cognitive (Spatial) Tasks

Level 1	Understand the concept of "getting out"
Level 2	Understand the meaning of "in" and "out"
Level 3	Understand the concepts of "go in," "come out," and "under"
Level 4	Understand "inside," "outside," "underneath," and "on top of"
Level 5	Understand the meanings of "put it around" and "take it off"
Level 6	Besides what was learned before, understand one more meaning of "up"
Level 7	Besides what was learned before, understand one more meaning of "next to"
Level 8	Besides what was learned before, understand the meanings of "close to," "behind"

Figure B.6: The Timing of Cognitive Skill (Spatial) Tasks across Difficulty Levels



Note: Level 1: Underdtand the concept of "getting out", Level 8: Besides what learned before, understand the meanings of "close to", "behind."

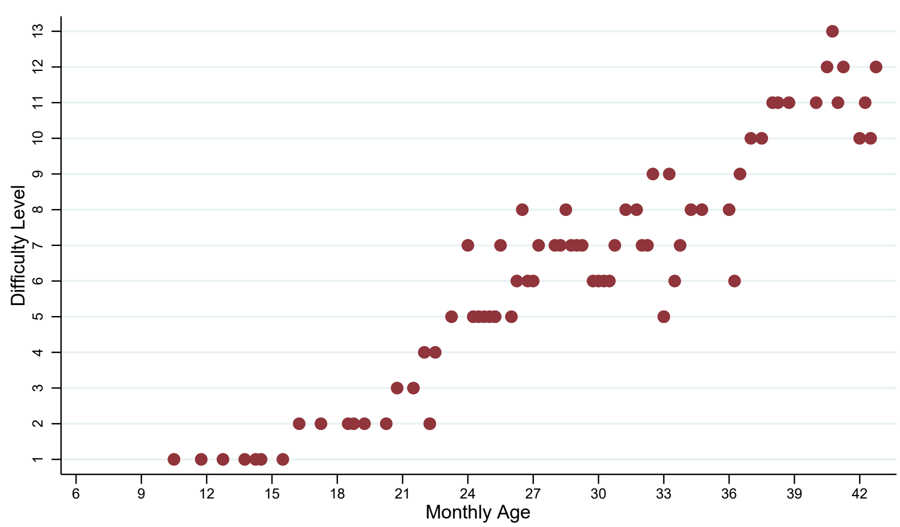
B.3.4.2 Knowing Objects and Objects' Functions

The knowing objects task set introduces preliteracy skills. It involves progressing interaction with pictures of objects and elements of storytelling. The tasks in the Cognitive Knowing Objects progress from a simple understanding of the concept of pictures by acknowledging with vocalizations, to using receptive (heard) language to identify certain pictures. Receptive language is a skill developed prior to an expressive language where a child forms words to communicate. The children must use their expressive language to complete the following tasks that increase with difficulty as they must develop more and more language to identify an increasing number of images. To progress through level 7 and beyond, the child must display an increasingly sophisticated understanding of the stories presented, first simply naming actions, then answering questions, then talking abstractly about the story. Levels 10, 11, 12, and 13 ask the child to take the information presented and build on it by discussing the uses of objects presented and making connections with other images.

Table B.8: Difficulty Level List for Cognitive (Understanding Objects) Tasks

Level 1	The child can look at the pictures and vocalize
Level 2	Name the objects and ask the child to point to the pictures accordingly
Level 3	The child can name the objects in one picture, and point to the named picture
Level 4	The child can name the objects in two or more pictures, and point to the named picture
Level 5	The child can point out named pictures, and say names of three or more
Level 6	The child can point out the picture mentioned, and correctly name the name of 6 or more pictures
Level 7	The child can talk about the pictures, answer questions, understand or names the verbs (eat, play, etc.)
Level 8	The child can follow the storyline, name actions and answer question
Level 9	The child can understand stories, and talk about the content in the pictures
Level 10	The child can keep up with the development of story
Level 11	The child can say the name of each graphic, discuss the role of each item, and then link the graphics in the card together
Level 12	The child can name the items in the picture, link the different pictures together, and discuss some of the activities in the pictures
Level 13	The child can name the things in the picture and talk about the function of objects

Figure B.7: The Timing of Cognitive Skill (Understanding Objects) Tasks across Difficulty Levels



Note: Level 1: Look at the pictures and vocalize; Level 13: The child can name the things in the picture and talk about the function of objects.

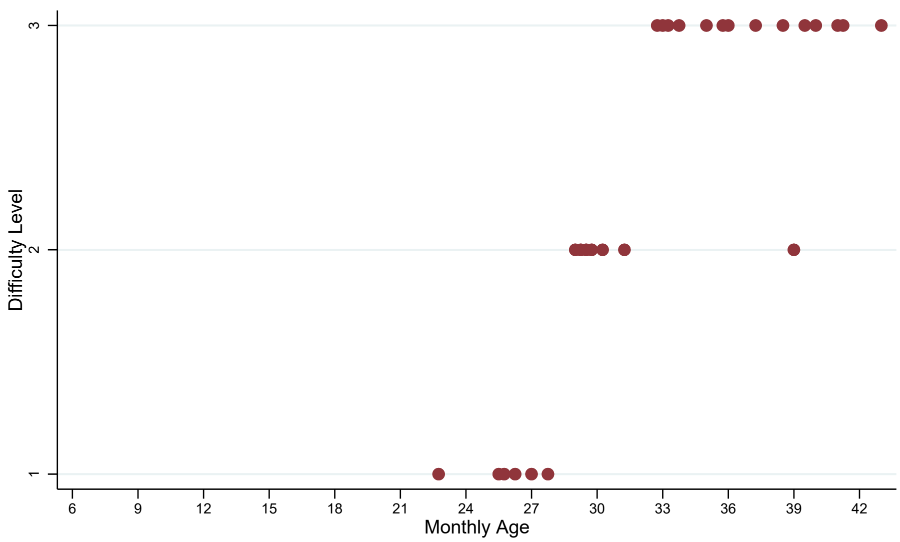
B.3.4.3 Color

In the color skill set, tasks progress from passive interactions (child hearing about color) to actively naming colors, to finally making connections with colors.

Table B.9: Difficulty Level List for Cognitive (Color) Tasks

Level 1	Caregiver talks about the color
Level 2	The child can identify the color
Level 3	Match different colors

Figure B.8: The Timing of Cognitive Skill (Understanding Color) Tasks across Difficulty Levels



Note: Level 1: Caregiver talks about color; Level 3: Understand color and match color with objects.

Table B.10: Difficulty Level List for Cognitive (Order: Understanding Upward, Forward, First, Some, All, Next, and Last) Tasks

Level 1	Child learns how to string beads and understands the meanings of "upward" and "downward"
Level 2	Understand the meanings of "upward," "downward," "first," and "then"
Level 3	Understand the concepts of "first," "finally," "in front of," and "behind"

Cognitive ability progresses into more abstract concepts of direction "upward" and "downward." Then, relative concepts of "first," "last," or "behind" are introduced.

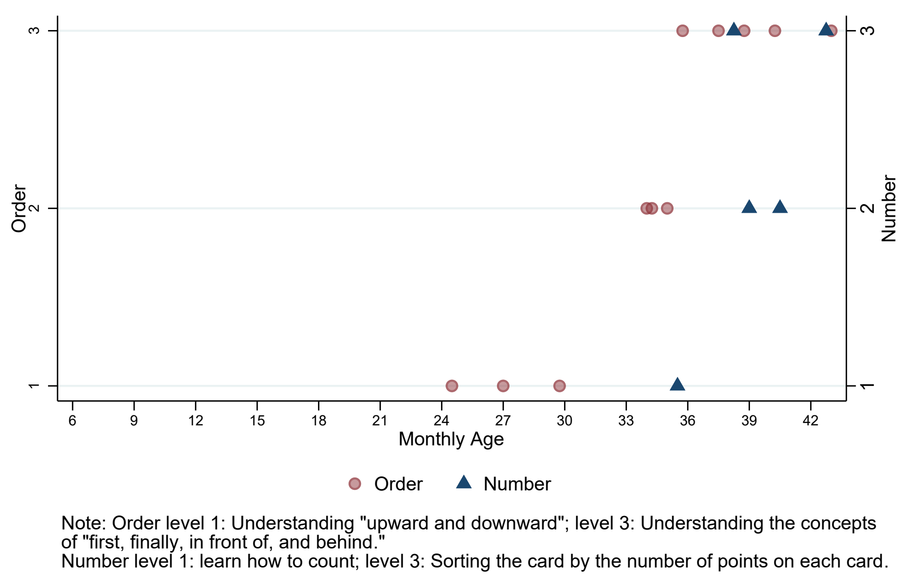
B.3.4.4 Number

Table B.11: Difficulty Level List for Cognitive (Number) Tasks

Level 1	Child learns how to count, can count up to 4
Level 2	Counting from 1 to 4, and then count two objects: 1, 2
Level 3	Children can count from 1 to 4 and sort the card by the number of points on each card

Number tasks progress from the learning of numbers in order to understanding one-to-one relationships of numbers to objects when counting. Finally, the concept of number representation is introduced.

Figure B.9: The Timing of Cognitive Skill (Understanding Order and Numbers) Tasks across Difficulty Levels



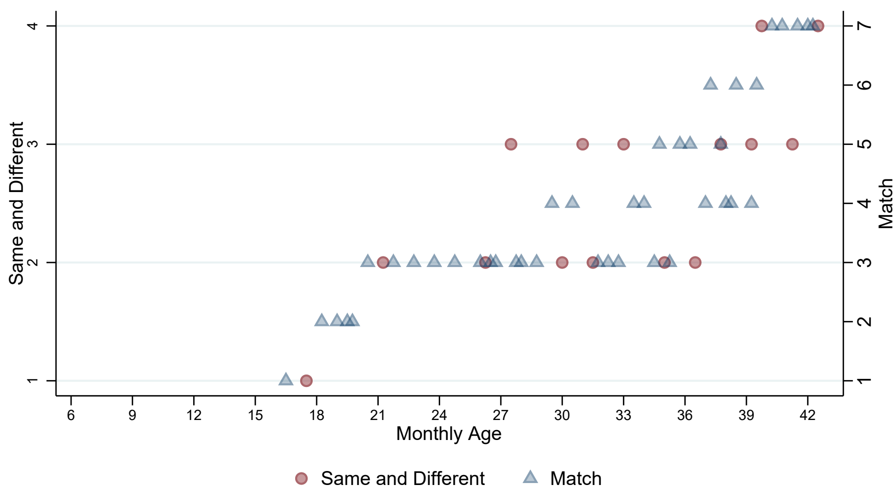
B.3.4.5 Match

These tasks consist of matching different pieces from simple puzzles to complicated puzzles. This set of tasks builds on the child's spatial awareness skills. The ability to fill in missing objects and understand how objects fit together is important in developing spatial awareness. The individual tasks progress from simply placing 1-2 puzzle pieces, completing the puzzle, making patterns, and using emerging language skills to describe pieces. As the children gain proficiency in these skills, they can complete puzzles of increasing complexity and restore the jumbled pieces to the original puzzle.

Table B.12: Difficulty Level List for Cognitive (Match) Tasks

Level 1	Put one piece into the puzzle
Level 2	The child is able to put at least two pieces in the puzzle
Level 3	The child can complete the simple puzzle
Level 4	The child can complete the puzzle and name different pieces
Level 5	The child learns to put together puzzle pieces to form the complete pattern
Level 6	With the caregiver's help, the child can complete the puzzle with more pieces
Level 7	The child can restore the puzzle to the original

Figure B.10: The Timing of Cognitive Skill (Matching and Understanding) Tasks across Difficulty Levels



Note: Same and Different level 1: teaching the meaning of "same", level 4 the child understands the concept of "same" and "different."
Match level 1: teaching one shape puzzle; level 7: match different shape pieces.

B.3.5 Language Skill

Language skill is the ability of children to communicate their needs, thoughts, feelings and ideas in a way that the caregiver can understand. It includes vocalizations, gestures, spoken words, and other signals.

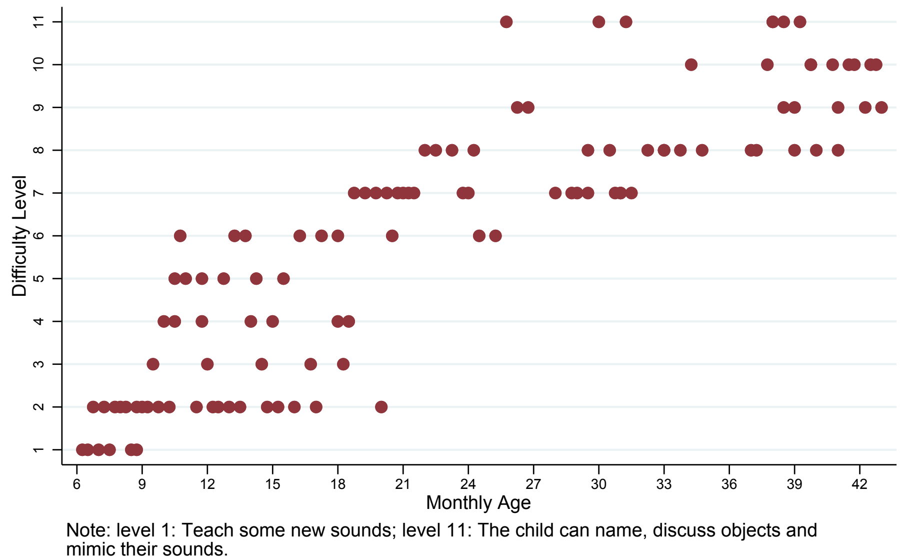
B.3.5.1 Learn words

Table B.13: Difficulty Level List for Language (Knowing Objects and Understanding Their Functions) Tasks

Level 1	Caregiver and baby make sounds to each other to interact
Level 2	Caregiver tells baby the things she does in the house
Level 3	To teach baby to recognize people's names
Level 4	Baby learns movements that show intimacy: clapping, bye-bye, and thank you
Level 5	Caregiver and child look at the pictures together, and let the child vocalize and touch the pictures
Level 6	Baby is to recognize at least one body part
Level 7	The child identifies and/or names ordinary objects
Level 8	The child points to the pictures which are being named, names one or more pictures, mimic the sound of the objects
Level 9	The child points to the pictures which are being named, names two or more pictures, mimic the sound of the objects
Level 10	The child points at 7 or more than 7 pictures and talk about them
Level 11	Teach the child some simple descriptive words and the child names objects at home, and tells the usage of those objects

The language skill tasks increase in difficulty with the expectation that the child will learn to identify and use expressive language to indicate understanding. The tasks begin with the baby passively listening as the caregiver makes sounds and speaks. The child then plays a more active role, expected to indicate understanding (receptive language) and use simple gestures to indicate meaning. The language skills tasks begin simply with the baby passively listening as the caregiver makes sounds and speaks. The child then plays a more active role, expected to indicate understanding (receptive language) and use simple gestures to indicate meaning. As understanding and vocabulary increase, the child will name more pictures and learn to describe them. Finally, the child will learn the names and uses of objects in the child's everyday environment.

Figure B.11: The Timing of Language Skill (Knowing Objects) Tasks across Difficulty Levels



B.3.5.2 Dialogue

In this set of tasks, the caregiver talks to the children.

Table B.14: Difficulty Level List for Language (Dialogue) Tasks

Level 1	Caregiver talks to the baby when doing housework
Level 2	Use words that child learned to answer or create a new conversation

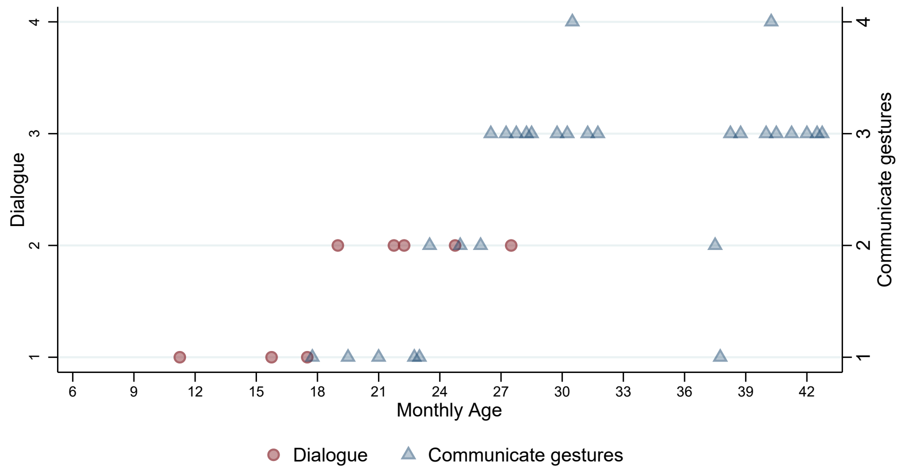
As the child grows, the caregiver progresses from simply narrating events to building on words the child has learned to scaffold language development.

B.3.5.3 Communicate Gestures

Table B.15: Difficulty Level List for Language (Communicate Gestures) Tasks

Level 1	The baby listens to simple instructions given by the caregiver
Level 2	Caregiver performs some activities with the child
Level 3	Let the child learn to talk about the pictures, act according to the pictures, answer questions, and name related actions

Figure B.12: The Timing of Language Skill (Communicate Gestures) Tasks across Difficulty Levels



Note: Dialogue level 1: Talking to the child; level 2: Using the words the child learned to create conversation.
Communicate gestures level 1: the child listens to simple instructions; and level 4: the child can act as other roles, e.g., father, mother.

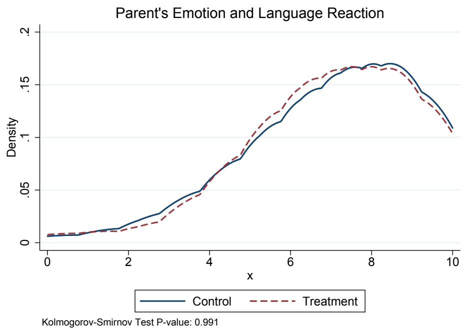
C Baseline Comparisons

In order to examine the quality of the randomization, in this section, we compare both the targeted and untargeted moments or distributions in the randomization design between the treatment and control groups. In Figure C.1, we give a comparison of the variables which are used in designing the matched pair. We can find that the control group and the treatment group have very similar distributions for the variables used for the randomization design. The Kolmogorov-Smirnov test p-values are all above 0.7, which indicates that we cannot reject the hypothesis that the distributions of the control and treatment groups are identical.

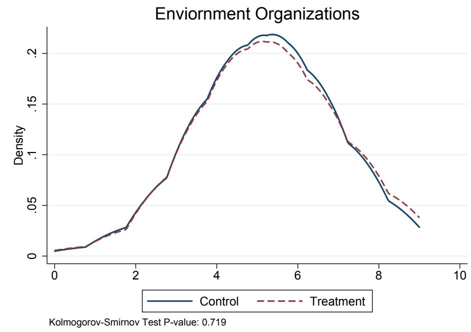
Next, we compare the variables which are not considered in the randomization estimation. We conduct this comparison for living conditions; family education levels, family structure, and economic conditions; pregnancy knowledge, pregnancy behavior, and the situations in pregnancy; children's health and development measures; and parent-child interaction.

Figure C.2 shows that well or spring water is the main water source for cooking in both control and treatment groups. About 95% of households have stable electricity for daily life. In Huachi county, there is a kind of traditional cave dwelling housing (Yaodong). 70% of households are still living in this kind of traditional housing. Figure C.3 shows the outward appearances of Yaodong in Huachi county. The fractions of different types of durable goods owned by each household are presented in Figure C.4. All t-test p-values are above 0.05, and there is no significant difference between treatment and control group households in terms of the ownership of durable goods. Almost every household has a cell phone and television. One notable fact is that about 70% of households have at least a motorcycle and above 20% of households own their cars. The ownership of an automobile is higher than 14%, which is the car parc rate in China. Most residents are living in a mountainous area. Cars or motorcycles are important tools to connect to places outside of the village.

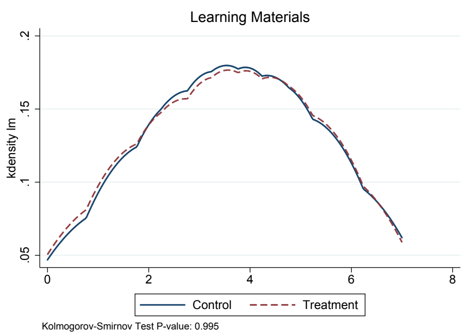
In Huachi county, the family structure is quite stable. For example, in Figure C.5 more than 98% of children's fathers and mothers are married or cohabitating. Figure C.6 shows the education distribution for different household members. The children's fathers and mothers have higher education levels than the grandfather-mother generation. More than 62% of children's fathers finished at least nine years of compulsory education. About 19% of the fathers graduate from high school or above. For the children's mothers, about 55% of them finished at least nine years of mandatory education, and about 11% are high school graduates



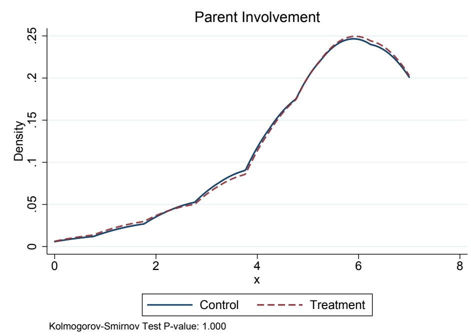
(a)



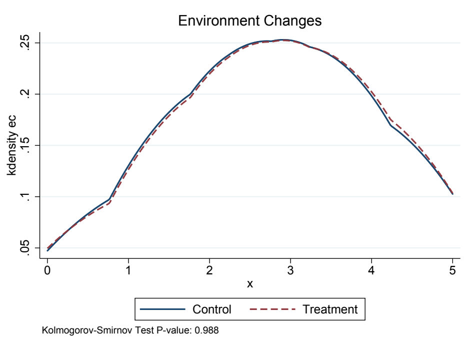
(b)



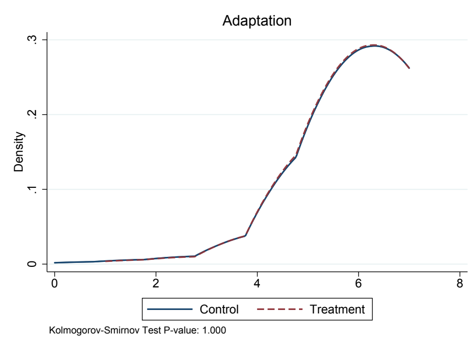
(c)



(d)

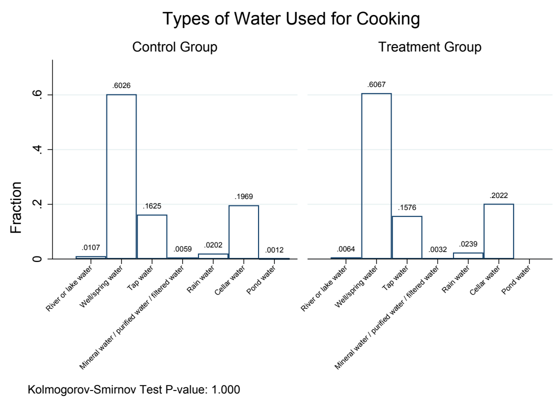


(e)

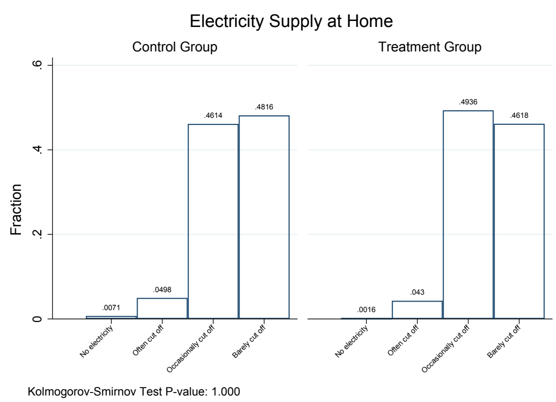


(f)

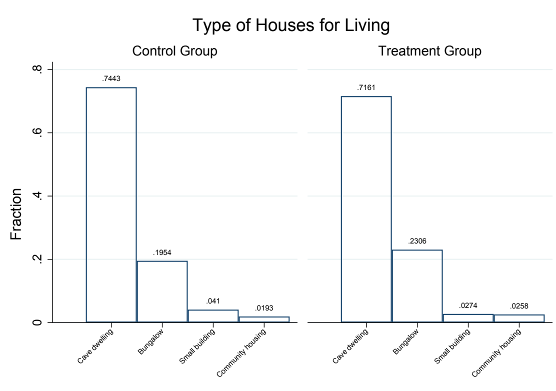
Figure C.1: Distributions of Outcomes Used in Designing Matched Village Pairs



(a)



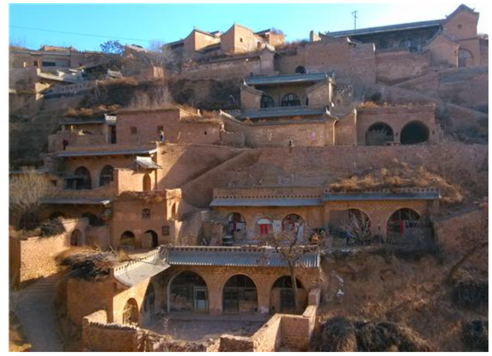
(b)



(c)

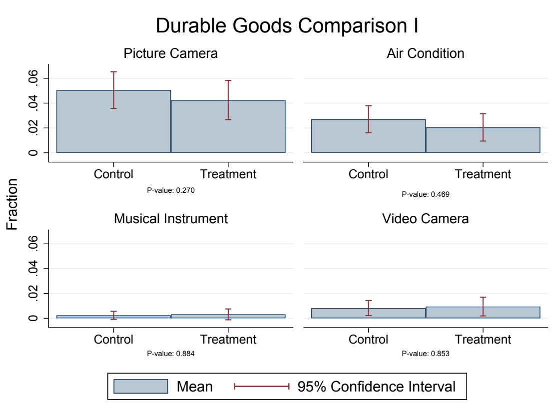
Figure C.2: Living Conditions

(a)

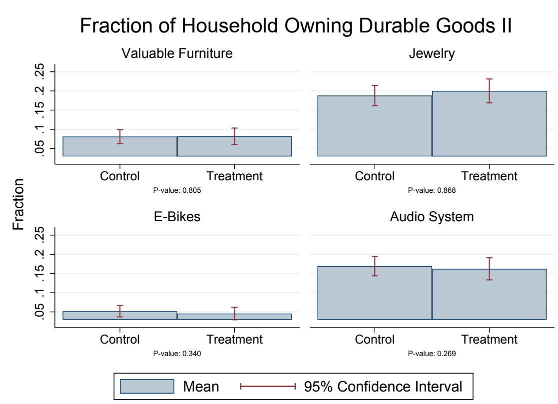


(b)

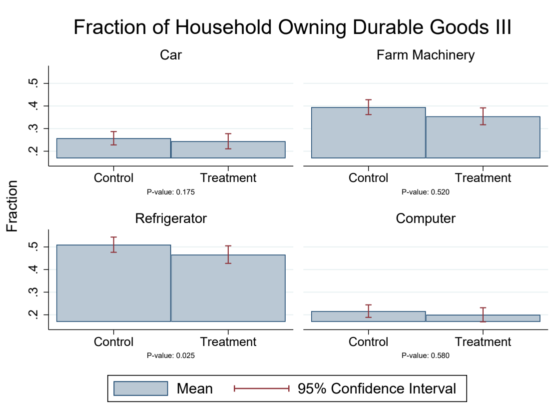
Figure C.3: Yaodong in Huachi county



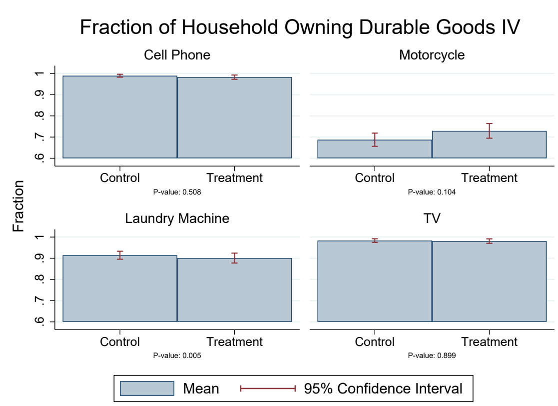
(a)



(b)



(c)



(d)

Figure C.4: Fraction of Households Owning Durable Goods

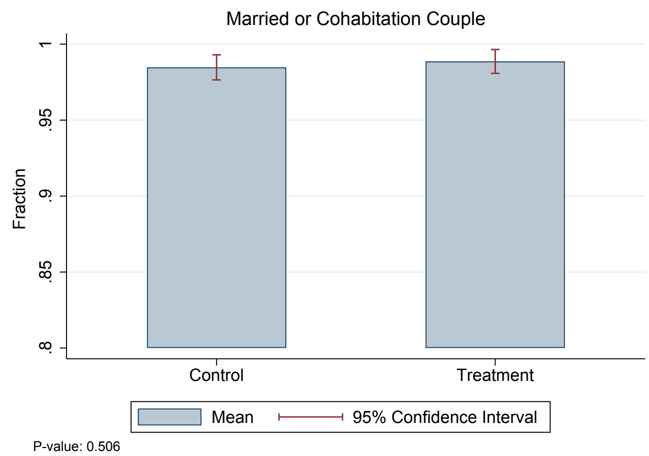


Figure C.5: Family Structure

or above. The fractions of mothers who have graduated primary school or middle school are similar to the fractions of the fathers. For the grandfather-mother generation, it is clear that grandmothers are less educated. More than 40% of them do not have any formal education. In general, we can see that the education distributions for different household members between the control and treatment groups are very close. The Kolmogorov-Smirnov test shows that the education level distributions are identical between the control and treatment groups.

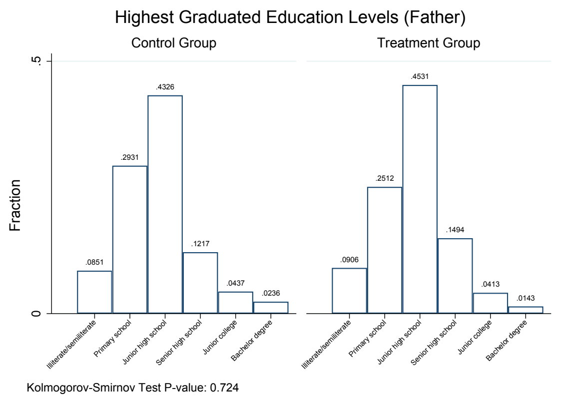
In Table C.1, we compare household annual income and consumption by categories. Since there are multiple income sources (e.g., wage income, agriculture income, and government subsidy) for the rural households, in the table we lay out main income sources of the control and treatment groups at the baseline. The column of "*p*-value" gives the statistics testing whether the mean values are different between the two groups. All *p*-values are greater than 5%, which means that we cannot reject the null hypothesis that the mean values of the two groups are equal to each other.

Figures C.7-C.9 give a summary of the knowledge of pregnancy and the performance in pregnancy. Figure C.7 provides the comparison of pregnancy knowledge between the control and treatment groups. In general, control group individuals have greater knowledge of pregnancy but we cannot reject the null that both groups have equal knowledge of pregnancy. Figure C.8 shows the pregnancy behaviors which would affect child health outcomes. We cannot find significant difference between the control and treatment groups. Almost no mother smokes or drinks during her pregnancy. More than 80% of the mothers had prenatal check experiences, and 60% of the mothers had prenatal checks in the first three months

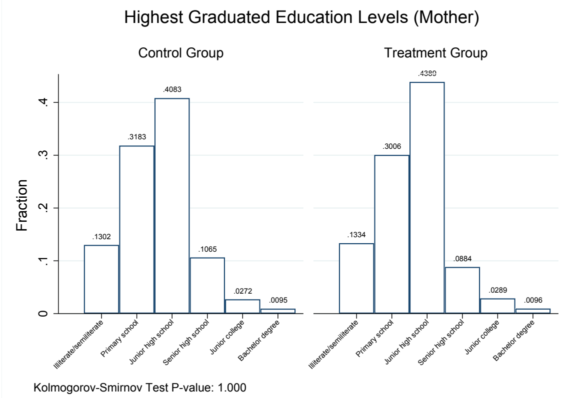
Table C.1: Consumption and Income Comparisons (Baseline)

Household Level	Control	Treatment	p-value
Agricultural Income	10284.66	7055.83	0.09
Standard Error	(1655.21)	(1912.39)	
Observations	704	515	
Government Subsidy	2780.42	2321.70	0.06
Standard Error	(220.42)	(244.40)	
Observations	751	567	
Remittance	15632.60	15969.94	0.87
Standard Error	(1233.56)	(2127.09)	
Observations	544	408	
Wage Income (After Tax)	34934.14	31255.06	0.39
Standard Error	(3102.52)	(4276.55)	
Observations	92	64	
Food Consumption	7861.47	9638.88	0.62
Standard Error	(1267.94)	(1956.49)	
Observations	703	513	
Total Consumption	42767.85	41796.35	0.92
Standard Error	(6504.43)	(10021.38)	
Observations	846	629	

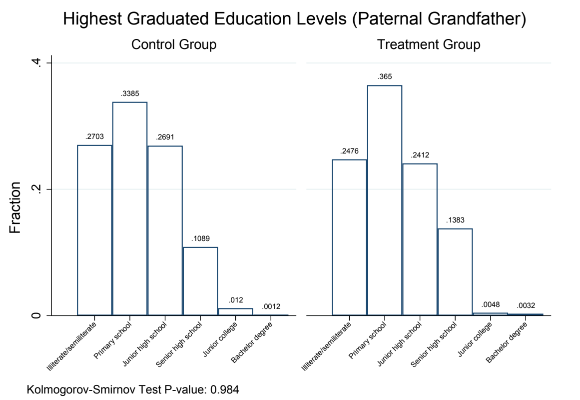
p-value is calculated by bootstrapping and clustering at the level of the randomized paired villages.



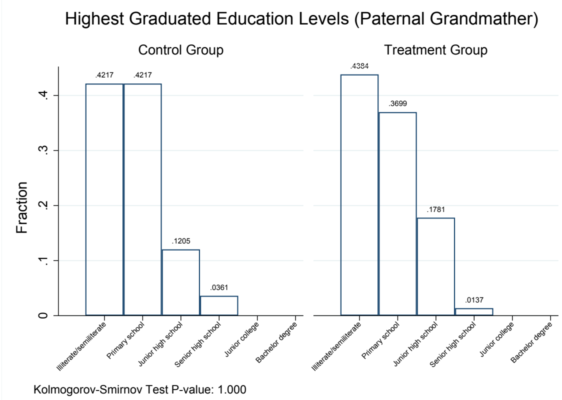
(a)



(b)

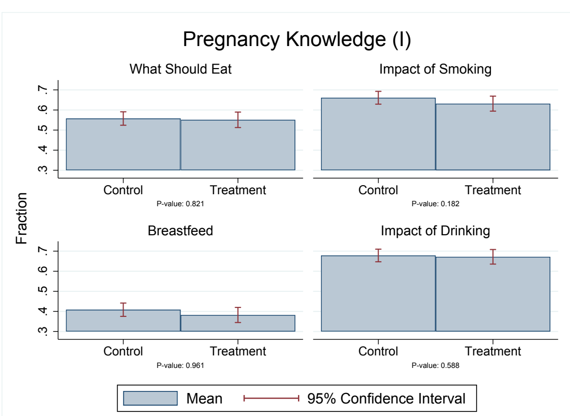


(c)

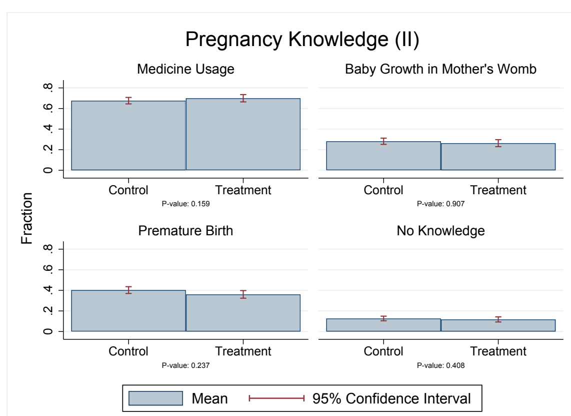


(d)

Figure C.6: Family Member Education Levels

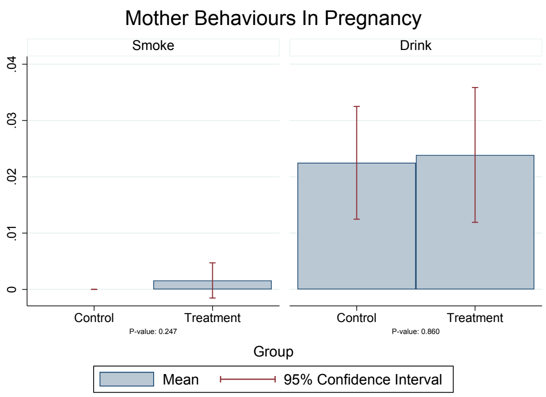


(a)

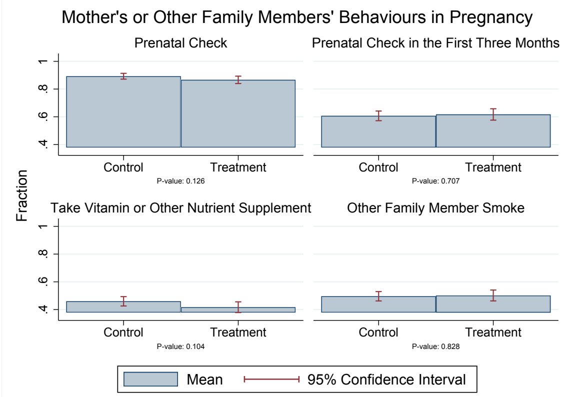


(b)

Figure C.7: Pregnancy Knowledge



(a)



(b)

Figure C.8: Pregnancy Behavior

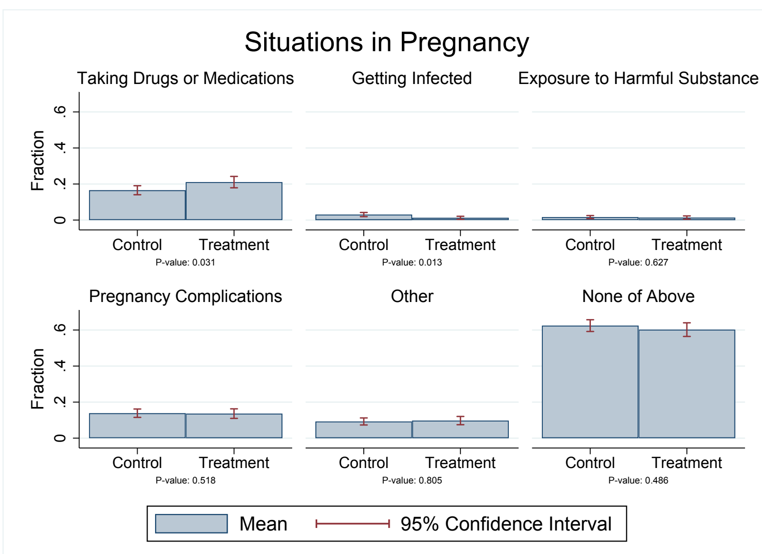
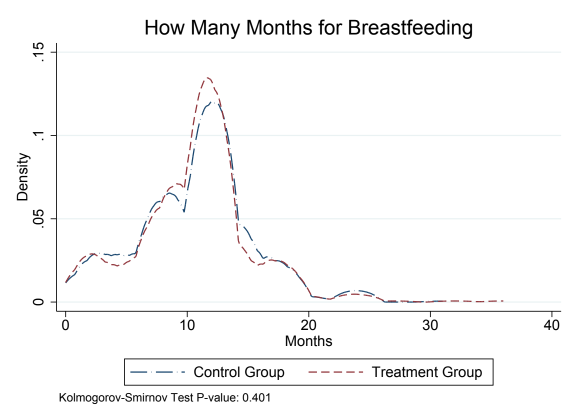
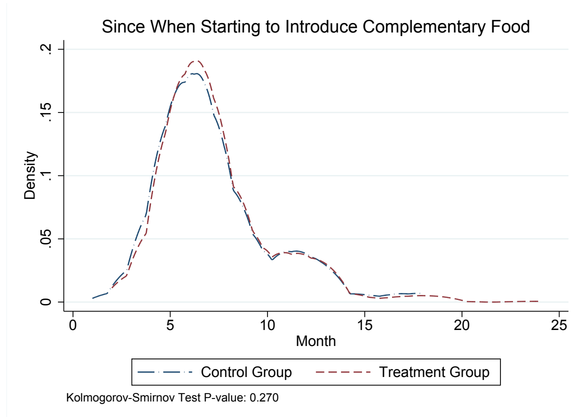


Figure C.9: Situations in Pregnancy

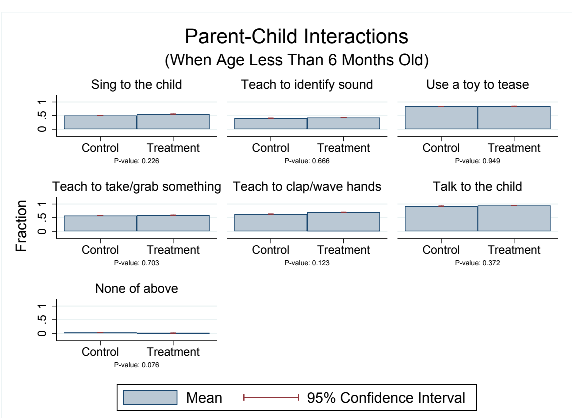


(a)

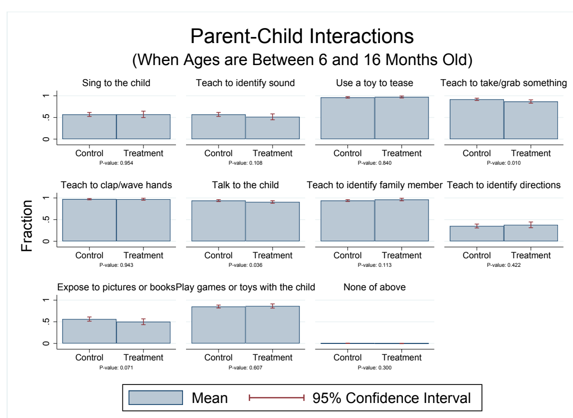


(b)

Figure C.10: Breastfeeding Behavior



(a)



(b)

Figure C.11: Parent-Child Interaction

of pregnancy. Also, we cannot find significant differences in the health conditions during pregnancy between the control and treatment groups in Figure C.9.

Table C.2 shows the comparison of Child Development Measures between the control and treatment groups at the baseline. Although we have fewer than 150 children's Denver test information at the baseline, for these children, we find that there is no significant difference between the two groups for different types of skills. Also, Table C.2 documents that the birth weight and height are very close between the two groups.

Figure C.10 gives the duration of breastfeeding and the time of introducing complementary food to the infants. The distributions are very close for both the control and treatment groups. From Figure C.11, we can find that the changes in parent-child interactions are related to the children's age. Also, there are no significant differences in both parent-child interactions between the two groups.

From the above comparisons, there are no significant differences in either target outcomes or the non-target variables between the control and treatment groups. In general, the randomization design works well in selecting matched pair villages.

D Denver II Test

In Figures D.1 and D.2, we present the Denver test implemented during the intervention for both English and Chinese versions. The items are over 99% consistent between the two versions.

Table C.2: Child Development Measures (Baseline)

	Control	Treatment	p-value
Denver Test			
Personal/Social	7.27	7.15	0.52
Standard Error	(0.12)	(0.18)	
Fine Motor Adaptive	12.27	12.23	0.81
Standard Error	(0.11)	(0.16)	
Language	11.14	11.08	0.75
Standard Error	(0.13)	(0.18)	
Gross Motor	11.56	11.09	0.02
Standard Error	(0.15)	(0.20)	
Total	42.24	41.56	0.17
Standard Error	(0.37)	(0.49)	
Observations	63	73	
Birth Weight	3.18	3.23	0.03
Standard Error	(.01)	(.02)	
Observations	765	687	
Birth Height	49.83	49.95	0.31
Standard Error	(0.08)	(0.12)	
Observations	675	643	
Monthly age at the baseline	11.39	9.63	0.00
Standard Error	(0.18)	(0.25)	
Observations	852	726	

p-value is calculated by bootstrapping at the level of the randomized paired villages.

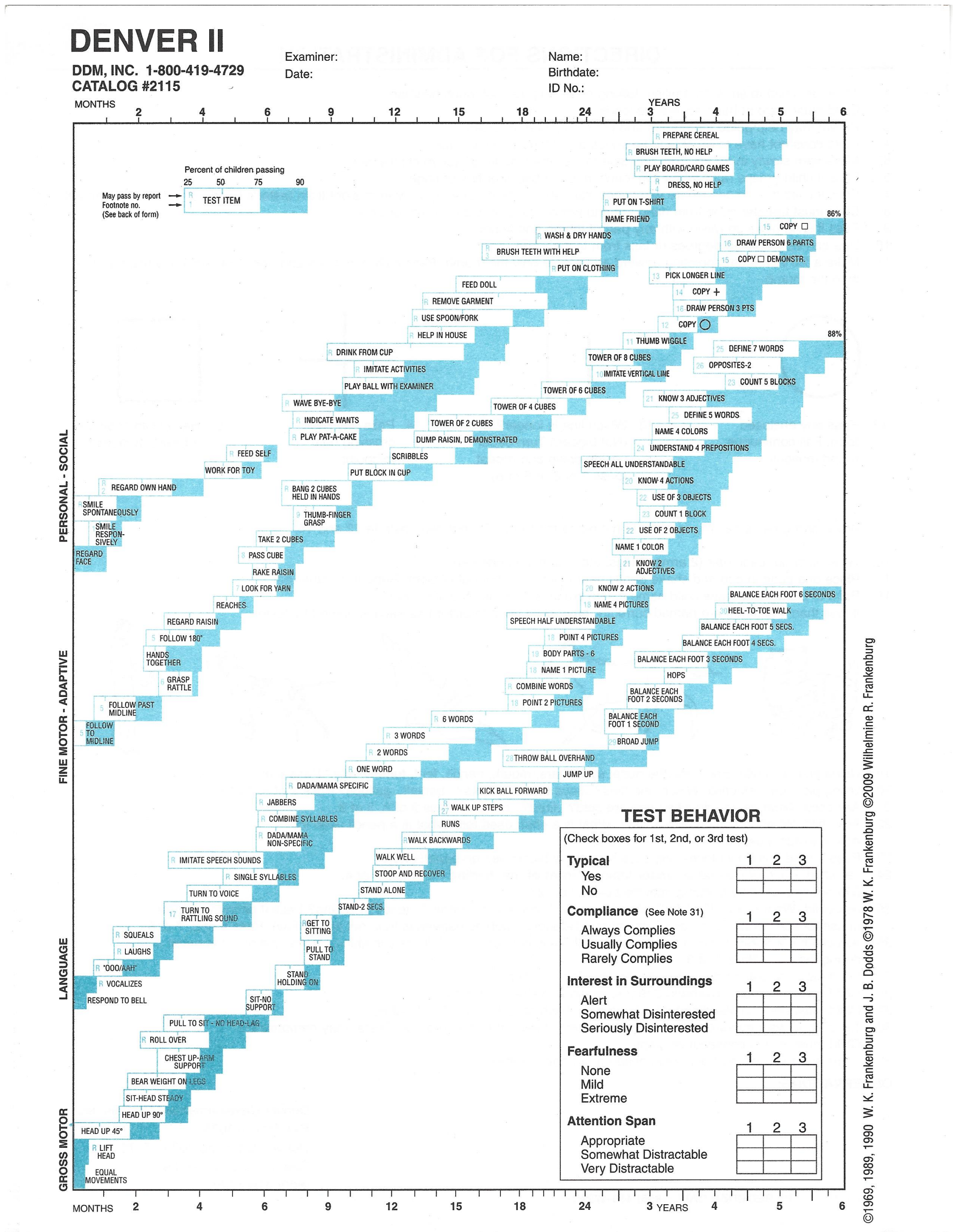


Figure D.1: English Denver Test

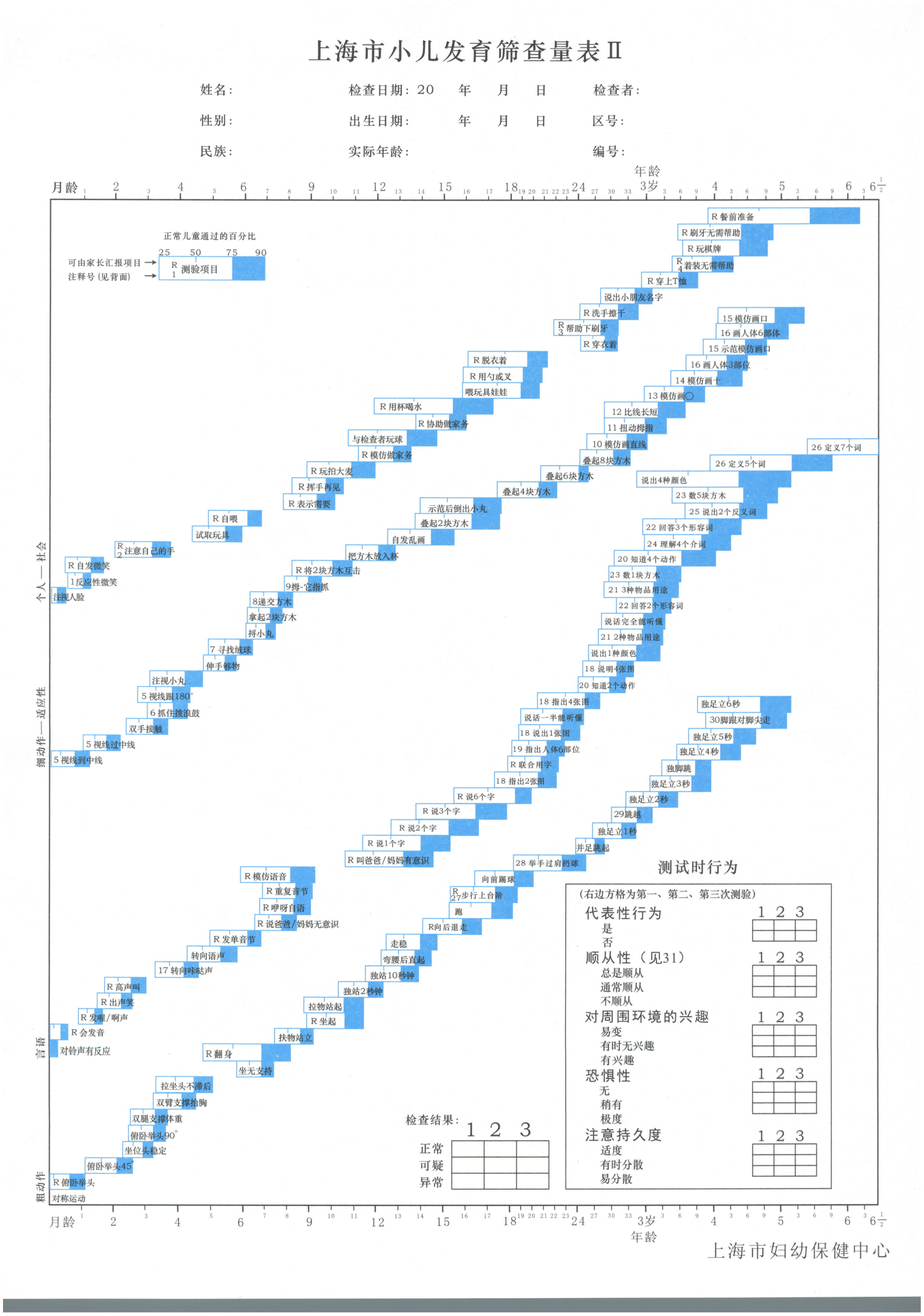


Figure D.2: Chinese Denver Test

E Linear Model Estimates on Raw Scores

Table E.1: Treatment Effects on Total Scores

	(1) All	(2) All	(3) Children $\leq$ 2 Yrs at Enrollment	(4) All	(5) Children $\leq$ 2 Yrs at Enrollment
			Midline		
Language and Cognitive	0.533***	0.569***	0.689***	0.634***	0.754***
	[0.162, 0.895]	[0.161, 0.969]	[0.299, 1.090]	[0.234, 1.036]	[0.347, 1.173]
Fine Motor	0.064	0.166	0.200	0.195	0.228
	[-0.104, 0.233]	[-0.075, 0.412]	[-0.033, 0.444]	[-0.052, 0.467]	[-0.014, .488]
Social-Emotional	0.206**	0.274***	0.259***	0.285***	0.271***
	[0.044, 0.372]	[0.094, 0.452]	[0.073, 0.453]	[0.115, 0.463]	[0.067, 0.477]
Gross Motor	-0.140	-0.121	-0.009	-0.119	-0.031
	[-0.391, 0.110]	[-0.398, 0.156]	[-0.277, 0.276]	[-0.391, 0.148]	[-0.295, 0.243]
			Endline		
Language and Cognitive	1.031***	0.966***	1.172***	1.041***	1.247***
	[0.599, 1.472]	[0.509, 1.427]	[0.735, 1.591]	[0.601, 1.489]	[0.813, 1.687]
Fine Motor	0.224	0.205	0.232**	0.238*	0.265**
	[-0.006, 0.457]	[-0.021, 0.424]	[0.023, 0.454]	[0.009, 0.472]	[0.019, 0.530]
Social-Emotional	-0.133	-0.159	-0.093	-0.136	-0.066
	[-0.299, 0.038]	[-0.342, 0.022]	[-0.260, 0.071]	[-0.319, 0.051]	[-0.240, 0.109]
Gross Motor	0.085	0.106	0.101	0.112	0.122
	[-0.244, 0.422]	[-0.184, 0.405]	[-0.190, 0.396]	[-0.174, 0.401]	[-0.166, 0.405]
Pre-treatment Covariates	No	No	No	Yes	Yes
IPW	No	Yes	Yes	Yes	Yes

1. The 95% confidence intervals in parentheses are constructed by wild bootstrap clustered at the village level.

2. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table E.2: Treatment Effects on Raw Scores

(Female)

	(1) All	(2) All	(3) Children $\leq$ 2 Yrs at Enrollment	(4) All	(5) Children $\leq$ 2 Yrs at Enrollment
			Midline		
Language and Cognitive	0.368	0.365	0.479**	0.388	0.518**
	[-0.089, 0.813]	[-0.105, 0.842]	[0.020, 0.939]	[-0.073, 0.882]	[0.100, 0.996]
Fine Motor	0.135	0.132	0.183	0.156	0.192
	[-0.110, 0.379]	[-0.113, 0.369]	[-0.0337, 0.396]	[-0.103, 0.418]	[-0.030, 0.411]
Social-Emotional	0.352**	0.348**	0.368***	0.382***	0.399***
	[0.098, 0.617]	[0.101, 0.584]	[0.125, 0.609]	[0.126, 0.640]	[0.132, 0.663]
Gross Motor	-0.043	-0.078	-0.034	-0.069	-0.055
	[-0.363, 0.284]	[-0.421, 0.276]	[-0.372, 0.298]	[-0.427, 0.303]	[-0.409, 0.318]
			Endline		
Language and Cognitive	0.775*	0.827*	0.859*	0.869*	0.894*
	[-0.118, 1.611]	[0.022, 1.618]	[0.007, 1.683]	[0.004, 1.728]	[0.073, 1.749]
Fine Motor	0.350	0.313	0.348	0.347	0.369
	[-0.044, 0.790]	[-0.043, 0.705]	[-0.004, 0.698]	[-0.066, 0.797]	[-0.040, 0.807]
Social-Emotional	-0.147	-0.169	-0.146	-0.164	-0.148
	[-0.339, 0.045]	[-0.363, 0.033]	[-0.363, 0.066]	[-0.333, 0.011]	[-0.325, 0.036]
Gross Motor	0.167	0.209	0.236	0.208	0.223
	[-0.434, 0.724]	[-0.346, 0.779]	[-0.347, 0.804]	[-0.321, 0.728]	[-0.319, 0.739]

1. The 95% confidence intervals in parentheses are constructed by wild bootstrap clustered at the village level.

2. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table E.3: Treatment Effects on Raw Scores

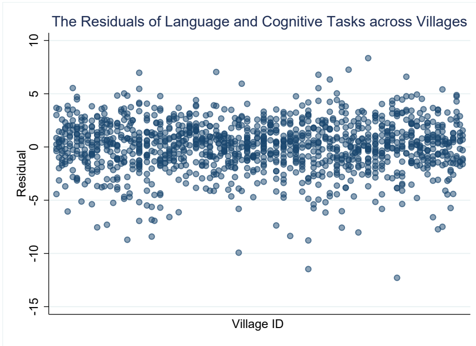
(Male)

	(1) All	(2) All	(3) Children ≤ 2 Yrs at Enrollment	(4) All	(5) Children ≤ 2 Yrs at Enrollment
			Midline		
Language and Cognitive	0.716***	0.807***	0.905***	0.853***	0.917***
	[0.204, 1.250]	[0.245, 1.378]	[0.354, 1.467]	[.306, 1.388]	[0.297, 1.520]
Fine Motor	0.046	0.193	0.208	0.211	0.231
	[-0.198, 0.304]	[-0.135, 0.551]	[-0.101, 0.539]	[-0.117, 0.590]	[-0.096, 0.604]
Social-Emotional	0.162	0.240	0.212	0.229	0.188
	[-0.083, 0.411]	[-0.012, 0.499]	[-0.071, 0.485]	[-0.044, 0.500]	[-0.101, 0.480]
Gross Motor	-0.154	-0.086	0.041	-0.126	-0.020
	[-0.437, 0.123]	[-0.447, 0.271]	[-0.277, 0.368]	[-0.426, 0.193]	[-0.306, 0.289]
			Endline		
Language and Cognitive	1.198***	0.948**	1.273***	1.037***	1.376***
	[0.548, 1.822]	[0.233, 1.635]	[0.724, 1.837]	[0.375, 1.730]	[0.766, 1.980]
Fine Motor	0.138	0.111	0.128	0.108	0.124
	[-0.118, 0.395]	[-0.151, 0.391]	[-0.132, 0.386]	[-0.146, 0.368]	[-0.136, 0.398]
Social-Emotional	-0.146	-0.194	-0.095	-0.181	-0.089
	[-0.391, 0.115]	[-0.456, 0.065]	[-0.315, 0.144]	[-0.479, 0.121]	[-0.369, 0.216]
Gross Motor	-0.060	-0.067	-0.066	-0.077	-0.059
	[-0.289, 0.171]	[-0.295, 0.168]	[-0.297, 0.156]	[-0.283, 0.134]	[-0.269, 0.161]

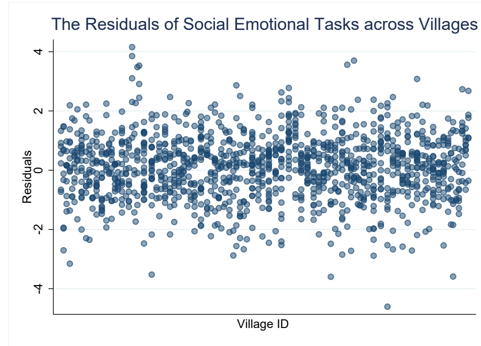
1. The 95% confidence intervals are constructed by wild bootstrap clustered at the village level.

2. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

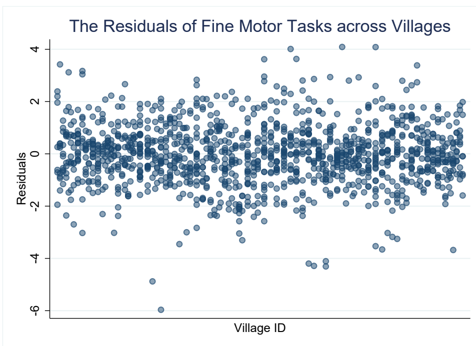
Figure E.1 shows that the residuals from these regressions are at best weakly correlated across villages.



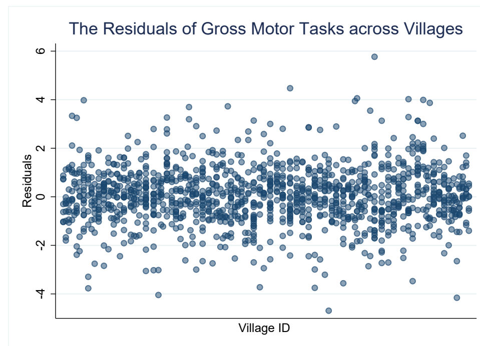
(a)



(b)



(c)



(d)

Figure E.1: Test of Residual Independence across Villages

F Wild Bootstrap Procedure

The Cameron, Gelbach, and Miller (2008) procedure:

1. From OLS estimation on the original sample, we obtain the estimates $\hat{\beta}$ and the CRVE $\hat{\Omega}$. Also, based on the null hypothesis $\boldsymbol{a'\beta} = \mathbf{0}$, we reestimate the model to obtain restricted estimates $\tilde{\beta}$ and residuals $\tilde{\boldsymbol{u}}$, calculate the cluster robust t statistic t_o

2. Do B iterations of this step. On the bth iteration:

   (a) Form a sample of V clusters $(\boldsymbol{\hat{y}_1^*, X_1}, \cdots, \boldsymbol{\hat{y}_V^*, X_V})$ by the following method. For each cluster v, $\boldsymbol{u_v^{*b}} = w_v^{*b}\boldsymbol{\tilde{u}_v}$ and the w_v^{*b} are independent realizations of an auxiliary random variable w^* with zero mean and unit variance; then form $\boldsymbol{y^{*b}} = \boldsymbol{X'\tilde{\beta}} + \boldsymbol{u_v^{*b}}$

   (b) Calculate the bootstrap estimates $\boldsymbol{\hat{\beta}^{*}}^{\boldsymbol{b}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y^{*b}}$ and the bootstrap covariance matrix and the bootstrap t-statistic t^{*b}

3. Reject H_0 at the level α if and only if $t_o < t^*_{[\alpha/2]}$ or $t_o > t^*_{[1-\alpha/2]}$

G Data, Attrition, and Nonresponse

This section documents data collection procedures, data attrition problems, and how we address data attrition problems. In January 2015, CDRF collected baseline information in Huachi county; 1,566 children were presented at that time. The RCT design was conducted based on the 1,566 children's survey information and village level administrative data, in which 796 children are in the treated villages and 770 children are in the control villages.

In September 2015, the home visiting intervention started. There was an eight-month gap between the baseline data collection and the first home visit interventions. The local field team made two modifications to the original protocol before they started the first home visit. The first modification is that they included 76 new children in the intervention who were not surveyed in January 2015 but were eligible in September 2015 and, in addition, they excluded most of the children who were older than two years old in September 2015 (about 150 children).[7] The second modification is that they excluded children with urban hukou (around 90). Therefore, after the two adjustments, in September 2015, the sample size was 1395, including 634 children in the treatment group and 761 children in the control group. In January 2016, 180 children from the younger cohort were added: 89 in the treatment group and 91 in the control group.[8] The data we use in our analysis include 1567 observations: 1395 of which were tracked since January 2015, plus 172 children for whom the baseline was January 2016 (8 children are missing from 180 samples). Finally, the sample we use includes 1,567 children, of which 715 are in the treatment group and 852 are in the control group.[9]

Table G.1 summarizes the sample created before the first intervention. The two main modifications targeted only the treatment group, hence most modifications came from the treatment group children. Since there were newly enrolled and also excluded children during this process, we examine the baseline comparison (the final sample with 1567 children) between the control and treated group children in Section C, and find that there is no significant difference between treatment and control groups. When the field group started the home visits, there were 715 children on their name list. Among these 715 children, 705 children

[7]For these 76 children, the field team collected their baseline information at the midline annual evaluation based on the parents retrospective responses.

[8]For these new children, they are a younger cohort from Huachi county and also take the Denver test assessment.

[9]Here, the reason why the final sample is 1,567 and not 1,566 is that we find two children share the same ID in the Denver test. We also find both children's information in the weekly home visit records. Therefore, we include the additional child.

participated in the home visits (i.e., compliance rate is above 98%).

Table G.1: Huachi County Data Sample Before the Intervention

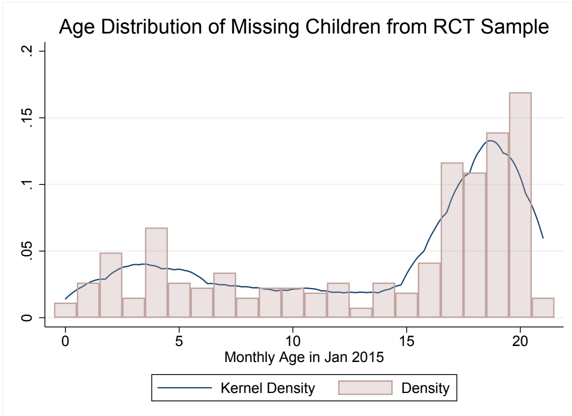
(January 2015-January 2016)

	Total	Treatment	Control
Baseline (January 2015)	1566	796	770
Adjustment in September 2015			
Adjustment 1 (include the children not surveyed in January 2015)	76	76	0
Adjustment 2 (exclude old and urban hukou children)	-247	-238	-9
New Enrollment in January 2016	180	89	91
Missing	-8	-8	0
Final Sample	1567	715	852

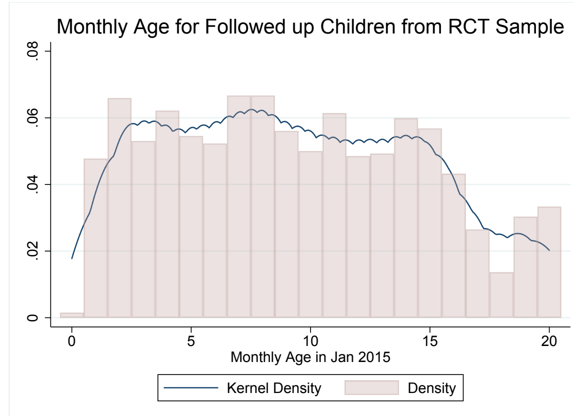
To have robust estimates, in the estimation, we also account for the sample adjustment in September 2015. According to the timeline, there were three stages at which data attrition occurred. The first stage is data attrition before February 2016. Before February 2016, 1,822 children were enrolled in the program, which included 1566 from January 2015, 76 from September 2015, and 180 from January 2016. After January 2016, 245 children were not followed up with. For the home visiting intervention, the final list was based on the 715 children in the treatment group previously discussed, of whom 705 had at least one home visit (close to 99% of children in the treatment group had been treated). By July 2017, the average number of home visits was 74.

Since both the midline and endline child development assessments were conducted in a short time window (e.g., two weeks), and the annual review assessments were conducted in the town center hospital, data attrition appeared at both rounds of assessments. Table G.1 shows the decomposition of the 1567 children in the followed-up sample. At the midline, there were 1301 children who attended the Denver test (i.e., 636 in the treatment group, and 633 in the control group); for family survey information, information on more than 1,430 children was recorded. At the endline, there were 1,073 children who attended the Denver test examination (i.e., 529 in the treatment group and 544 in the control group). The family survey was conducted for 1,189 children (i.e., 569 in the treatment group and 620 in the control group).[10]

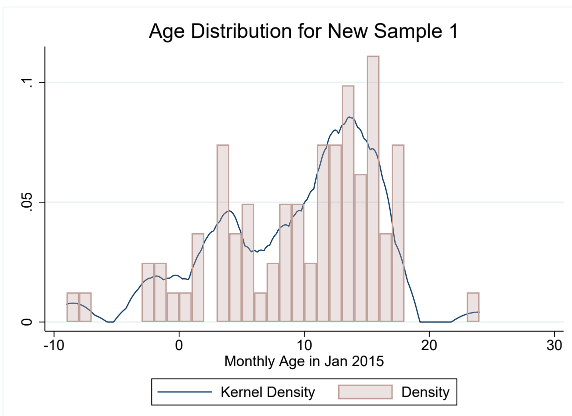
[10] The data missing in the two year annual evaluations can be treated as independent events (e.g., there is no evidence to show being missing from the July 2016 evaluation is correlated with being missing from July 2017 evaluation).



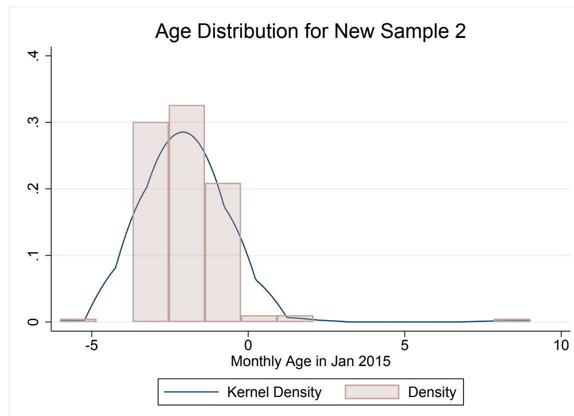
(a)



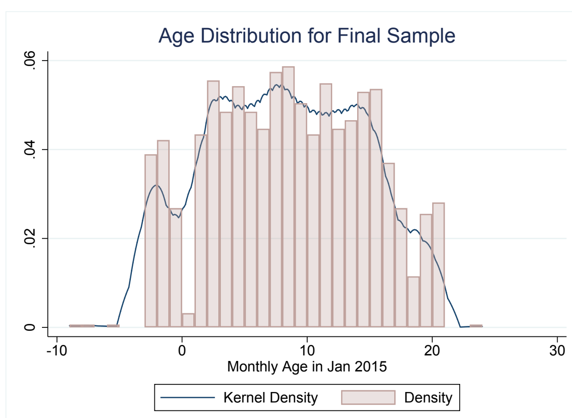
(b)



(c)



(d)



(e)

Figure G.1: Age Distribution for the Samples before February 2016

We use the inverse probability weighting method to address missing data problems (Tsiatis, 2006). The IPW estimator solves the following minimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(\boldsymbol{z_i})} \right) (y_i - \boldsymbol{x_i'\beta})^2$$

where s_i is the indicator if we can use observation i, $P(\boldsymbol{z_i})$ is the propensity score of observation i being observed. $\boldsymbol{z_i}$ is a vector of baseline variables which are always observed for everyone. Here, we mainly focus on the estimation of propensity scores at different stages. We present the distribution of the propensity score of the samples with missing data in Figure G.2 and compare them with the distribution in the observed outcome samples. In the figures, we show the distribution of the probabilities of being missing at three stages for the observed samples and missing data samples. Our model performs very well in predicting missingness: for observed outcomes samples, the probability of missing is low. For most of them the probability is less than 0.2. For the samples with missing outcome data, the propensity scores are close to being uniformly distributed. For both rounds, the estimated propensity scores are far away from 1 which means we do not need to trim the data. We thus avoid the inconsistency due to data trimming (Maasoumi and Wang, 2019).

Table G.2: Propensity Score for Missing Data

	First Stage	Second Stage		Third Stage	
	Miss	Miss Denver	Miss Survey	Miss Denver	Miss Survey
	Before January 2016	At Midline	At Midline	At Endline	At Endline
Older Than 24 Months	1.5674***	0.0568	0.0654	0.2484+	0.1042
(In September 2015)	(0.2163)	(0.1491)	(0.1548)	(0.1361)	(0.1300)
Monthly Age	0.0341*	0.0135	0.0151+	-0.0084	0.0012
(In September 2015)	(0.0160)	(0.0087)	(0.0092)	(0.0076)	(0.0072)
The population of the Village	0.0011	-0.0004	-0.0003	0.0003	0.0005+
	(0.0008)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Number of Households in the Village	-0.0001	-0.0000	0.0002	0.0001	0.0004
	(0.0010)	(0.0004)	(0.0005)	(0.0004)	(0.0004)
Size of Working Population	-0.0042**	0.0004	0.0006	-0.0010+	-0.0008
	(0.0016)	(0.0006)	(0.0006)	(0.0005)	(0.0005)
Poor Village or Not	1.3732**	0.1078	0.2037	-0.0161	0.0833
	(0.5066)	(0.2058)	(0.2129)	(0.1833)	(0.1702)
Number of Persons Receiving Social Welfare	-0.0051**	0.0006	-0.0005	0.0025***	0.0017**
	(0.0017)	(0.0008)	(0.0009)	(0.0007)	(0.0007)
Mean Years of Schooling among Villagers	-0.5857*	0.1179	0.1485	-0.1169	-0.1114
	(0.2714)	(0.1230)	(0.1218)	(0.1083)	(0.1020)
Fraction of Interviewed Children Who Are Left-Behind	1.1623	-0.4473	0.7090	-1.0534	-1.3794+
	(2.6825)	(0.9716)	(0.9408)	(0.8324)	(0.7728)
HOME - Sum of Warmth/Responsiveness Items	0.0253	0.0247	0.0153	0.0049	-0.0063
	(0.0410)	(0.0271)	(0.0284)	(0.0239)	(0.0226)
HOME - Sum of Verbal Skills Items	0.1647+	0.0661	0.0992	-0.0273	-0.0643
	(0.0908)	(0.0587)	(0.0626)	(0.0498)	(0.0478)
HOME - Sum of Harshness/Discipline Items	-0.0201	0.0399	0.0448	-0.2120*	-0.1247+
	(0.1206)	(0.0863)	(0.0882)	(0.0876)	(0.0778)
HOME - Sum of Stimulation/Teaching Items	0.0046	-0.0182	-0.0206	-0.0179	-0.0060
	(0.0263)	(0.0181)	(0.0189)	(0.0161)	(0.0154)
HOME - Dum of Outings Items	0.1590**	0.0305	0.0323	-0.0356	-0.0311
	(0.0614)	(0.0430)	(0.0448)	(0.0383)	(0.0362)
Fraction of Children Taking Nutrition Package	-0.3803*	-0.2581*	-0.1977+	-0.0393	-0.1036
	(0.1533)	(0.1011)	(0.1063)	(0.0887)	(0.0842)
Fraction of Children Taking Nutrition Package without Interruption	-0.5557*	-0.0612	-0.0765	0.0259	0.0342
	(0.2332)	(0.1340)	(0.1401)	(0.1144)	(0.1091)
Number of Eligible Kids Living at Home in Interviewed Households in This Village	0.8225***	-0.0958*	-0.0681	-0.0574	-0.0614+
	(0.1450)	(0.0416)	(0.0433)	(0.0353)	(0.0337)
Fraction of Parents Willing to Participate in This Village	0.3958	0.0042	-0.0963	-0.1754	-0.0485
	(0.3759)	(0.2008)	(0.2029)	(0.1720)	(0.1707)
Number of Eligible Kids in Households That Would Be Willing to Participate	-0.7809***	0.0992*	0.0723	0.0592	0.0508
	(0.1464)	(0.0454)	(0.0471)	(0.0381)	(0.0364)
Fraction of Interviewed Households Planning to Migrate with the Child	0.1123	0.1415	0.1828	0.1495	0.0961
	(0.2381)	(0.1565)	(0.1613)	(0.1415)	(0.1374)
Distance between Home Visitor's Home and the Village	0.2559***	0.0126	0.0071	-0.0013	-0.0046
	(0.0666)	(0.0105)	(0.0057)	(0.0063)	(0.0064)
Whether Living in Chengguan Village	0.6431**	-0.6649*	-0.9191**	0.1176	0.1962
	(0.1975)	(0.2681)	(0.3190)	(0.1898)	(0.1809)
Family Migrate out of County	2.5438***	1.6588*	1.7525**		
	(0.3410)	(0.6442)	(0.6755)		
Refuse Home Visit	2.8099***	-5.0985	-5.1383		
	(0.3545)	(167.8907)	(166.4867)		
Refuse Home Visit in September 2015		6.3453	6.3040		
		(167.8891)	(166.4851)		
Constant	-4.1707	-2.2239*	-2.9191**	0.7335	0.9051
	(2.5367)	(1.0230)	(0.9568)	(0.9001)	(0.8452)
Observations	1823	1576	1576	1576	1576

Standard errors in parentheses

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

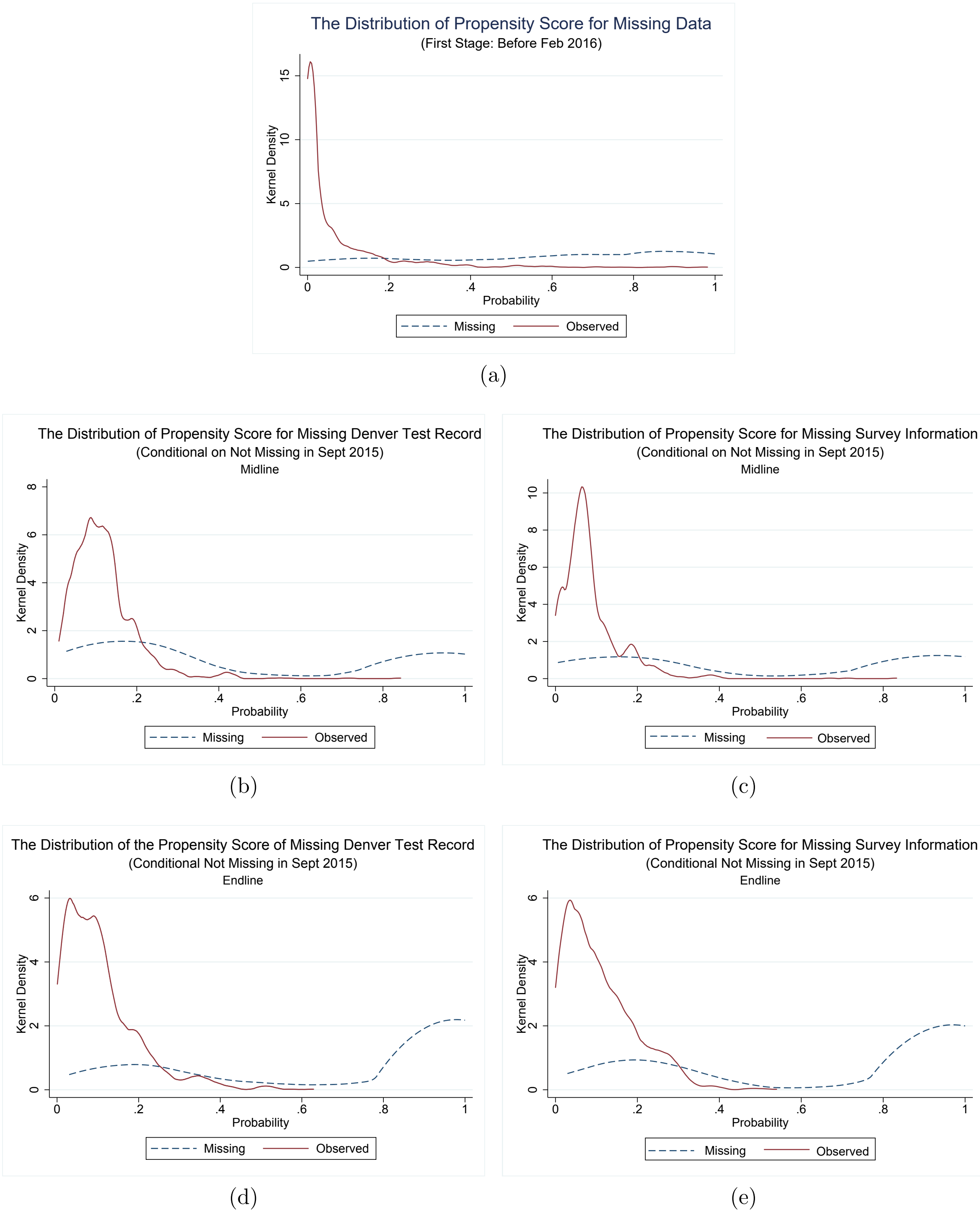


Figure G.2: Missing Data Propensity Score Distributions

H Measures of Interactions and Extracted Factors

H.1 Measures of Interactions

China REACH collects weekly records of child performance on lessons that measure child development during the weekly home visit intervention. The supervisors record home visitor, parent, and child interaction activities at least once per month, making it possible for us to examine their impacts. These measures are only recorded for the treatment group. We exploit variation within the treatment group, which, as we document below, is substantial.

We have detailed measures to evaluate the interaction quality between home visitors and caregivers and the visited children per visit. These observation-based measures were recorded by the program supervisors who randomly visited each household at least once per month at randomly selected times. During the home visit, the program supervisor evaluated the home visit's quality in three dimensions: the quality of the home visitor's teaching ability, the interaction quality between home visitor and caregiver; and the interaction quality between the home visitor and the child. We use the following measure as the teaching ability: "1. Does the home visitor bring the curriculum to the household? 2. Does the home visitor properly use the curriculum? 3. Has the home visitor prepared the home visit in advance? 4. Has the home visitor chosen the teaching materials and tasks which are suitable to the child's age? 5. Does the home visitor focus on language development? 6. During the home visit, what is the home visitor's attitude?" The measures we use for the interaction quality are listed in Table H.1.

Table H.1: Observed Interactions

Between Home Visitor and Caregiver
Has the home visitor explained the task content and lesson target to the caregiver?
Has the home visitor shown the lessons and given examples to the caregiver?
Does the home visitor ask the caregiver to play the lessons with the child alone?
Does the caregiver ask the home visitor about lessons in the next week?
Has the home visitor listened to the caregiver?
Has the home visitor answered the caregiver's questions?
Has the home visitor asked for the caregiver's opinions?
Does the home visitor encourage and help the caregiver?
Is the relationship between the home visitor and caregiver friendly, understandable, and cooperated?
Has the home visitor discuss with a caregiver or other persons about the content not related to the home visiting?
Between Home Visitor and Child
Has the home visitor shown the lessons and given examples to the child?
Has the home visitor explained the lesson to the child?
Does the home visitor listen to the child and respond to the child's voice or action?
Does the home visitor praise the child when the child tries to master one task?
Does the home visitor use language to communicate with the child when the child is completing the lessons?
Does the home visitor give the child enough time to explore the materials and finish the lessons?
Is the relationship between the home visitor, and the child friendly, understandable, and cooperative?

Note: The interaction quality measures are recorded by the supervisor of the program at least once per month.

Table H.2: Measure Variances

	Measure	Teaching Ability	Home Visitor-Caregiver Interaction	Home Visitor-Child Interaction
Midline	Standard Deviation (σ)	0.410	0.696	1.032
	Coefficient of Variation ($\frac{\sigma}{M}$)	0.324	0.260	0.189
Endline	Standard Deviation (σ)	0.400	0.701	1.002
	Coefficient of Variation ($\frac{\sigma}{M}$)	0.355	0.335	0.209

H.2 Factor Model of Interaction

As documented above, we have detailed measures on home visitor's teaching ability, the interaction measures between home visitor and caregiver (child). To summarize them, we estimate the latent factors of home visitor's teaching ability, and the interaction quality factors between home visitor and caregiver (child). We use a separate notation in this section.

Denote $M_{ia}^{j,l}$ as the measure j at household i at the child's age a and the γ_{ia}^{l} is the latent factors l represents different factors (i.e., teaching ability, the interaction quality between home visitor and caregiver, and the interaction quality between home visitor and child).

$$M_{ia}^{j,l} = X_{ia}'\beta + \alpha^{j}\gamma_{ia}^{l} + \epsilon_{ia}^{j,l} \tag{1}$$

We have estimated the factor model by MLE assuming normal errors. We estimate the latent factor l based on the empirical Bayes method: the empirical conditional posterior distribution of the latent factor is given as

$$g(\gamma^{l}|M^{l}, X; \beta, \alpha) = \frac{\mu(M^{l}|X, \gamma^{l}; \beta, \alpha, \phi(\gamma^{l}))}{\int \mu(M^{l}|X, \gamma^{l}; \beta, \alpha, \phi(\gamma^{l}))d\gamma^{l}} \tag{2}$$

Therefore, the latent factor estimates is given as $\hat{\gamma}^{l} = \int \gamma g(\gamma|M^{l}, X; \beta, \alpha)d\gamma^{l}$.

Table H.3: Treatment Effects on Standardized Language and Cognitive Scores

Variable	(1)	(2)	(3)	(4)	(5)
Denver Tasks	All	All	Children $\leq$ 2 Yrs at Enrollment	All	Children $\leq$ 2 Yrs at Enrollment
			Midline		
Treatment	0.384*	0.451*	0.468*	0.498*	0.497*
	[0.008,0.746]	[0.033,0.862]	[0.036,0.911]	[0.101,0.887]	[0.054,0.944]
Treatment × Interaction Between	0.379	0.509	0.451	0.559	0.495
Home Visitor and Caregiver	[-0.220,1.009]	[-0.117,1.199]	[-0.151,1.098]	[-0.033,1.154]	[-0.079,1.097]
Treatment × Interaction Between	0.223	0.337	0.265	0.291	0.235
Home Visitor and Child	[-0.200,0.634]	[-0.048,0.707]	[-0.119,0.644]	[-0.065,0.643]	[-0.127,0.577]
Treatment × Teaching Ability	1.028	0.585	0.622	0.631	0.649
	[-0.532,2.483]	[-0.976,2.06]	[-0.981,2.133]	[-0.804,2.028]	[-0.839,2.063]
			Endline		
Treatment	1.000***	0.924***	0.996***	1.018***	1.062***
	[0.590,1.405]	[0.498,1.356]	[0.613,1.375]	[0.580,1.458]	[0.666,1.451]
Treatment × Interaction Between	0.959	1.21*	1.051*	1.343**	1.231**
Home Visitor and Caregiver	[-0.056,2.051]	[0.254,2.282]	[0.252,1.857]	[0.357,2.416]	[0.386,2.098]
Treatment × Interaction Between	0.255	0.223	0.203	0.178	0.176
Home Visitor and Child	[-0.283,0.696]	[-0.273,0.622]	[-0.332,0.632]	[-0.357,0.626]	[-0.385,0.654]
Treatment × Teaching Ability	-0 .512	-0 .803	-0 .871	-0 .764	-0 .916
	[-2.725,1.63]	[-2.891,1.274]	[-3.146,1.297]	[-2.885,1.401]	[-3.176,1.482]
Pre-Treatment Covariates	No	No	No	Yes	Yes
IPW	No	Yes	Yes	Yes	Yes

Notes: 1. The 95% confidence intervals in brackets are constructed using the wild bootstrap clustered at the village level.

2. The mean and variance for the standardized score are estimated from the pooled sample of the control group children.

3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table H.4: Treatment Effects on Standardized Social-Emotional Scores

Variable Denver Tasks	(1) All	(2) All	(3) Children ≤ 2 Yrs at Enrollment	(4) All	(5) Children ≤ 2 Yrs at Enrollment
			Midline		
Treatment	0.509*	0.637**	0.339	0.617**	0.293
	[0.093,0.935]	[0.166,1.093]	[-0.080,0.750]	[0.203,1.031]	[-0.116,0.700]
Treatment × Interaction Between	0.968**	0.971*	1.192**	0.906*	1.181**
Home Visitor and Caregiver	[0.278,1.722]	[0.153,1.815]	[0.406,1.962]	[0.093,1.709]	[0.386,1.949]
Treatment × Interaction Between	0.664**	0.568**	0.554*	0.459*	0.442
Home Visitor and Child	[0.179,1.176]	[0.135,1.050]	[0.086,1.055]	[0.009,0.944]	[-0.050,0.977]
Treatment × Teaching Ability	-0 .930	-0 .785	-0 .845	-0 .538	-0 .636
	[-2.581,0.650]	[-2.263,0.687]	[-2.261,0.749]	[-2.030,1.047]	[-2.132,0.990]
			Endline		
Treatment	-.270	-.327	-.211	-.305	-.180
	[-0.675,0.148]	[-0.789,0.109]	[-0.622,0.195]	[-0.716,0.138]	[-0.567,0.226]
Treatment × Interaction Between	-0 .470	-0 .391	-0 .698	-0 .522	-0 .761
Home Visitor and Caregiver	[-1.355,0.574]	[-1.357,0.687]	[-1.505,0.134]	[-1.465,0.585]	[-1.493,0.032]
Treatment × Interaction Between	-0 .181	-0 .123	-0 .168	-0 .157	-0 .166
Home Visitor and Child	[-0.542,0.163]	[-0.450,0.226]	[-0.472,0.170]	[-0.451,0.162]	[-0.491,0.177]
Treatment × Teaching Ability	1.778*	1.400	1.643*	1.588*	1.675*
	[0.437,3.095]	[-0.108,2.847]	[0.231,3.027]	[0.150,2.973]	[0.146,3.024]
Pre-Treatment Covariates	No	No	No	Yes	Yes
IPW	No	Yes	Yes	Yes	Yes

Notes: 1. The 95% confidence intervals in brackets are constructed using the wild bootstrap clustered at the village level.

2. The mean and variance for the standardized score are estimated from the pooled sample of the control group children.

3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table H.5: Treatment Effects on Standardized Fine Motor Scores

Variable	(1)	(2)	(3)	(4)	(5)
Denver Tasks	All	All	Children ≤ 2 Yrs at Enrollment	All	Children ≤ 2 Yrs at Enrollment
			Midline		
Treatment	0.251	0.334	0.397	0.349	0.416
	[-0.246,0.762]	[-0.203,0.887]	[-0.116,0.903]	[-0.192,0.908]	[-0.081,0.932]
Treatment × Interaction Between	0.392	0.574	0.593	0.594	0.618
Home Visitor and Caregiver	[-0.357,1.209]	[-0.253,1.548]	[-0.200,1.465]	[-0.333,1.553]	[-0.252,1.506]
Treatment × Interaction Between	-0 .054	-0 .188	-0 .234	-0 .205	-0 .233
Home Visitor and Child	[-0.523,0.383]	[-0.781,0.385]	[-0.916,0.35]	[-0.760,0.344]	[-0.838,0.323]
Treatment × Teaching Ability	0.547	0.626	0.608	0.612	0.522
	[-1.295,2.479]	[-1.142,2.358]	[-1.195,2.456]	[-1.334,2.374]	[-1.363,2.408]
			Endline		
Treatment	0.661*	0.665*	0.665*	.733**	.700*
	[0.122,1.176]	[0.113,1.219]	[0.118,1.225]	[0.190,1.276]	[0.144,1.245]
Treatment × Interaction Between	0.348	0.435	0.314	0.764	0.695
Home Visitor and Caregiver	[-0.533,1.277]	[-0.460,1.419]	[-0.495,1.256]	[-0.111,1.727]	[-0.140,1.615]
Treatment × Interaction Between	-0 .247	-0 .205	-0 .236	-0 .256	-0 .291
Home Visitor and Child	[-0.758,0.195]	[-0.654,0.235]	[-0.719,0.251]	[-0.791,0.212]	[-0.851,0.207]
Treatment × Teaching Ability	-0 .604	-0 .826	-0 .797	-0 .961	-0 .971
	[-2.398,1.066]	[-2.581,0.806]	[-2.648,0.91]	[-2.699,0.696]	[-2.784,0.799]
Pre-Treatment Covariates	No	No	No	Yes	Yes
IPW	No	Yes	Yes	Yes	Yes

Notes: 1. The 95% confidence intervals in brackets are constructed using the wild bootstrap clustered at the village level.

2. The mean and variance for the standardized score are estimated from the pooled sample of the control group children.

3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table H.6: Treatment Effects on Standardized Gross Motor Scores

Variable Denver Tasks	(1) All	(2) All	(3) Children ≤ 2 Yrs at Enrollment	(4) All	(5) Children ≤ 2 Yrs at Enrollment
			Midline		
Treatment	-0.018	-0.041	-0.004	-0.083	-0.075
	[-0.581,0.552]	[-0.612,0.524]	[-0.636,0.649]	[-0.571,0.438]	[-0.626,0.46]
Treatment × Interaction Between	0.469	0.629	0.528	0.518	0.454
Home Visitor and Caregiver	[-0.439,1.277]	[-0.327,1.615]	[-0.553,1.539]	[-0.395,1.432]	[-0.551,1.421]
Treatment × Interaction Between	0.214	0.145	0.114	0.194	0.164
Home Visitor and Child	[-0.205,0.666]	[-0.343,0.627]	[-0.350,0.600]	[-0.321,0.705]	[-0.367,0.701]
Treatment × Teaching Ability	-0.518	-0.520	-0.422	-0.559	-0.508
	[-2.604,1.470]	[-2.606,1.489]	[-2.363,1.492]	[-2.228,1.254]	[-2.359,1.262]
			Endline		
Treatment	0.030	0.110	0.128	0.141	0.160
	[-0.582,0.638]	[-0.475,0.708]	[-0.456,0.734]	[-0.424,0.716]	[-0.374,0.722]
Treatment × Interaction Between	0.105	0.103	0.141	0.291	0.296
Home Visitor and Caregiver	[-1.357,1.597]	[-1.205,1.463]	[-1.066,1.337]	[-1.092,1.657]	[-0.980,1.504]
Treatment × Interaction Between	0.408	0.279	0.267	0.266	0.249
Home Visitor and Child	[-0.447,1.316]	[-0.574,1.213]	[-0.613,1.296]	[-0.661,1.317]	[-0.700,1.283]
Treatment × Teaching Ability	0.593	0.694	0.716	0.582	0.651
	[-1.884,3.083]	[-1.801,3.239]	[-2.029,3.32]	[-2.048,3.288]	[-2.071,3.204]
Pre-Treatment Covariates	No	No	No	Yes	Yes
IPW	No	Yes	Yes	Yes	Yes

Notes: 1. The 95% confidence intervals in brackets are constructed using the wild bootstrap clustered at the village level.

2. The mean and variance for the standardized score are estimated from the pooled sample of the control group children.

3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

I Robustness Check for Factor Normalization

Factor models require normalizations as long as we seek to separate factors from factor loadings. We use the items in the category "Baseline" to normalize the first four loadings for the four skills studied in this paper. We use a self-explanatory simplified notation. The items listed assign $\alpha^{jk} = 1$ to the items listed and zero otherwise, while the remaining factor loadings are freely specified. In Table I.2, we show means of the latent factor loadings under different normalizations and also test whether they are different from our original normalization estimates. We find that the results are quite stable across different normalization choices if we choose the normalized items in the medium difficulty level range.

Table I.1: The List of Normalized Task Items

	Original	(1)	(2)	(3)	(4)	(5)
Social-Emotional	Wash and Dry Hands	Wash and Dry Hands	Wash and Dry Hands	Wash and Dry Hands	Drink From Cup	Drink From Cup
Fine Motor	Imitate Vertical Line	Thumb Wiggle	Imitate Vertical Line	Imitate Vertical Line	Imitate Vertical Line	Thumb Wiggle
Language and Cognitive	Combine Words	Combine Words	Name Body Parts 6	Name Body Parts 6	Name Body Parts 6	Name Body Parts 6
Gross Motor	Broad Jump	Broad Jump	Broad Jump	Balance Each Foot 1 Second	Balance Each Foot 1 Second	Balance Each Foot 1 Second

Table I.2: Skill Loading Mean Comparison under Different Normalizations

Control							Treatment						
		Different Normalizations							Different Normalizations				
	Original	(1)	(2)	(3)	(4)	(5)		Original	(1)	(2)	(3)	(4)	(5)
	Language and Cognitive Skill Loadings							Language and Cognitive Skill Loadings					
Mean	0.442	0.453	0.442	0.462	0.439	0.454	Mean	0.656	0.697	0.697	0.725	0.721	0.709
S.D.	(0.371)	(0.320)	(0.346)	(0.354)	(0.380)	(0.381)	S.D.	(0.478)	(0.447)	(0.457)	(0.447)	(0.479)	(0.462)
p-value		0.335	0.969	0.125	0.876	0.599	p-value		0.008	0.001	0.000	0.037	0.076
	Gross Motor Skill Loadings							Gross Motor Skill Loadings					
Mean	0.712	0.710	0.716	0.741	0.737	0.768	Mean	0.668	0.639	0.690	0.690	0.681	0.692
S.D.	(0.422)	(0.425)	(0.388)	(0.373)	(0.419)	(0.451)	S.D.	(0.453)	(0.405)	(0.433)	(0.397)	(0.418)	(0.416)
p-value		0.897	0.690	0.088	0.345	0.069	p-value		0.092	0.013	0.120	0.529	0.299
	Fine Motor Skill Loadings							Fine Motor Skill Loadings					
Mean	0.437	0.431	0.465	0.429	0.451	0.470	Mean	0.539	0.543	0.565	0.556	0.539	0.541
S.D.	(0.269)	(0.248)	(0.263)	(0.249)	(0.300)	(0.292)	S.D.	(0.240)	(0.228)	(0.237)	(0.219)	(0.240)	(0.252)
p-value		0.348	0.005	0.606	0.438	0.082	p-value		0.672	0.097	0.390	0.989	0.944
	Social-Emotional Skill Loadings							Social-Emotional Skill Loadings					
Mean	0.259	0.245	0.238	0.227	0.220	0.214	Mean	0.223	0.177	0.190	0.188	0.168	0.171
S.D.	(0.275)	(0.279)	(0.270)	(0.278)	(0.314)	(0.338)	S.D.	(0.260)	(0.227)	(0.227)	(0.202)	(0.276)	(0.290)
p-value		0.119	0.005	0.059	0.138	0.118	p-value		0.000	0.001	0.020	0.037	0.046

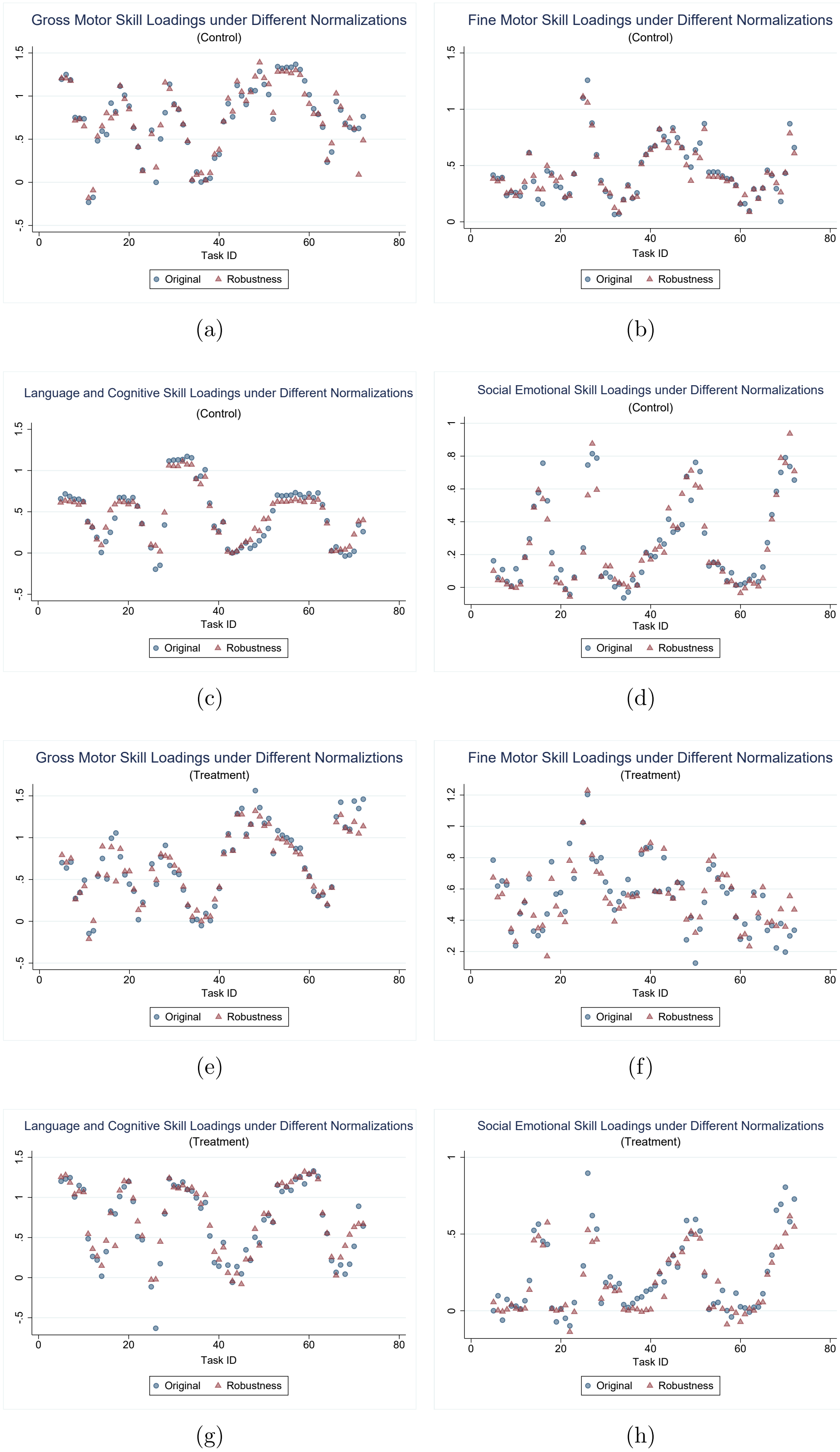


Figure I.1: The Comparison of Latent Skill Loadings under Different Normalizations

Table I.3: The Test of Equality in the Loadings between Treatment and Control ($\boldsymbol{\alpha}^{\boldsymbol{j,1}} = \boldsymbol{\alpha}^{\boldsymbol{j,0}}$)

	Social-Emotional Skills	Fine Motor Skills	Language and Cognitive Skills	Gross Motor Skills
	Step Down p-values			
	Social-Emotional Tasks			
Play Ball with Examiner	0.147	**0.042**	**0.028**	**0.032**
Help in House	0.193	**0.049**	**0.024**	0.131
Drink From Cup	0.054	**0.038**	**0.021**	0.059
Feed Doll	0.497	**0.021**	**0.031**	**0.019**
Use Spoon/fork	0.199	0.195	**0.023**	0.390
Remove Garment	0.121	0.239	**0.026**	**0.036**
Put on Clothing	0.152	**0.018**	0.066	**0.039**
Brush Teeth with Help	**0.027**	0.087	0.153	**0.045**
Name Friend	**0.047**	0.230	0.361	0.312
Put on T-shirt	0.519	0.437	0.390	**0.031**
Dress No Help	0.365	0.290	0.201	0.362
Play Board/Card Games	0.238	0.119	0.182	0.383
Brush Teeth no Help	0.476	0.436	0.511	0.433
	Language and Cognitive Tasks			
DaDa/MaMa Specific	0.336	**0.026**	0.098	**0.030**
One Word	0.162	**0.030**	0.270	0.176
Two Words	0.051	**0.025**	0.447	0.167
3 Words	0.051	**0.015**	0.130	**0.041**
6 Words	**0.033**	**0.013**	0.109	0.174
Point 2 Pictures	0.108	**0.012**	0.124	0.284
Body Parts 6	0.076	**0.034**	0.081	**0.045**
Name 1 Picture	0.431	**0.035**	0.109	0.073
Speech Half Understandable	0.061	**0.018**	0.181	0.061
Point 4 Pictures	0.454	0.112	0.510	0.091
Know 2 Actions	**0.048**	0.067	**0.046**	**0.035**
Name 4 Pictures	0.072	0.161	0.057	**0.049**
Name 1 Color	0.167	0.150	0.450	**0.037**
Use of 2 Objects	0.102	0.107	0.164	**0.043**
Speech all Understandable	0.059	0.387	0.416	0.060
Know 2 Adjectives	**0.045**	0.393	0.332	0.059
Use of 3 Objects	0.206	0.101	0.602	**0.018**
Count 1 Block	0.072	0.458	0.469	**0.041**
Know 4 Actions	0.153	0.463	0.510	0.063
Understand 4 Prepositions	0.084	0.216	0.253	0.135
Know 3 Adjectives	0.198	0.187	0.091	0.109
Opposites 2	0.318	0.211	0.060	0.303
Count 5 Blocks	0.298	0.253	0.407	**0.038**
Name 4 Colors	0.078	0.134	0.051	0.200

Table I.4: The Test of Equality in the Loadings between Treatment and Control ($\boldsymbol{\alpha}^{j,1} = \boldsymbol{\alpha}^{j,0}$)

	Social-Emotional Skills	Fine Motor Skills	Language and Cognitive Skills	Gross Motor Skills
	Step Down p-values			
Fine Motor Tasks				
Scribbles	0.162	**0.033**	**0.034**	**0.037**
Tower of 2 Cubes	0.058	0.060	**0.031**	0.429
Dump Raisin Demonstrated	0.131	**0.032**	**0.020**	0.090
Tower of 4 Cubes	0.428	**0.050**	**0.032**	0.073
Tower of 6 Cubes	0.146	**0.031**	0.248	**0.010**
Tower of 8 Cubes	0.274	**0.016**	0.219	**0.035**
Thumb Wiggle	0.173	0.613	0.668	**0.039**
Pick Longer Line	0.080	0.284	0.218	0.103
Copy Circle	**0.039**	0.371	0.481	**0.043**
Copy +	0.165	0.514	**0.026**	0.147
Draw Person 3 Parts	0.120	0.310	0.162	0.331
Gross Motor Tasks				
Get to Sitting	0.344	0.090	**0.048**	0.101
Pull to Stand	0.564	0.085	0.052	0.088
Stand 2 Seconds	0.607	0.087	**0.040**	0.060
Stand 10 Seconds	0.555	0.066	**0.036**	**0.047**
Stoop and Recover	0.180	0.052	**0.025**	0.236
Walk Well	0.093	**0.044**	**0.023**	0.051
Walk Backwards	0.075	0.087	**0.024**	0.421
Runs	0.370	**0.048**	**0.019**	0.445
Walk up Steps	0.383	**0.036**	**0.018**	0.052
Kick Ball Forward	0.128	**0.028**	**0.020**	0.254
Throw Ball Overhand	0.322	**0.019**	0.099	0.084
Jump up	0.239	**0.016**	0.095	0.087
Balance Each Foot 1 Second	0.203	**0.013**	0.239	0.396
Balance Each Foot 2 Second	0.241	0.342	0.249	0.375
Balance Each Foot 3 Second	0.045	0.158	**0.031**	0.044
Hops	0.152	0.521	0.073	0.161
Balance Each Foot 4 Second	0.674	0.133	**0.030**	**0.017**
Balance Each Foot 5 Second	0.395	0.509	0.159	**0.014**
Heel-to-toe Walk	0.688	0.358	0.076	0.347
Balance Each Foot 6 Seconds	0.477	0.101	0.059	0.351

J Consistency and Asymptotic Normality of Individual Factors (Wang, 2020) and a Factor Estimation Procedure from Chen, Fernández-Val, and Weidner (2021)

We use the analysis of Wang (2020), who proves the consistency and asymptotic normality of estimators of $\boldsymbol{\theta}_i^{j(k)}$ under conditions we satisfy. We actually know some of the factors he estimates, so our model is a special case of his. We apply an estimation procedure for factors proposed by Chen, Fernández-Val, and Weidner (2021).[11] We use a simplified self-explanatory notation in this appendix to facilitate exposition.

(1) For each iteration k, given the set of parameters $\{\boldsymbol{\beta}^{j(k)}, \boldsymbol{\theta}_i^{j(k)}, \boldsymbol{\alpha}^{j(k)}, \boldsymbol{\delta}^{j(k)}\}$, define $\mu_{ij}^j = \boldsymbol{X}^{j\prime}\boldsymbol{\beta}^{(k)} + \delta^{j(k)} + (\boldsymbol{\theta}_i^{j(k)\prime})\boldsymbol{\alpha}^{j(k)}$.

(2) **E-step**: Calculate

$$\begin{aligned}\hat{Y}_i^{j(k)} &= E(Y_{ij}^{j*}|Y_i^j, \boldsymbol{X}_i^j, \boldsymbol{\beta}^{j(k)}, \boldsymbol{\alpha}^{j(k)}, \delta^{j(k)}, \boldsymbol{\theta}_i^{j(k)}) \\ &= \mu_i^{j(k)} + (Y_i^j - \Phi^{j(k)}(\mu_i^{j(k)}))\phi^{j(k)}(\mu_i)/\{\Phi(\mu_i^{j(k)})(1-\Phi(\mu_i^{j(k)}))\}.\end{aligned}$$

(3) **M-step** conditional maximization steps:

- Update $\boldsymbol{\beta}^j$: $\boldsymbol{\beta}^{j(k+1)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\hat{Y}^{j(k)} - \delta^{j(k)} - (\boldsymbol{\theta}^{j(k)})'\boldsymbol{\alpha}^{j(k)})$.
- Update δ^j: $\delta^{j(k+1)} = \sum_i(\hat{Y}_i^{j(k)} - \boldsymbol{X}_i'\boldsymbol{\beta}^{j(k+1)} - (\boldsymbol{\theta}_i^{j(k)})'\boldsymbol{\alpha}^{j(k)})/N_{Ij}$.
- Update $\alpha_{j,m}^{j(k+1)}$, where m indicates the mth latent factor, since for each latent factor, we have one item m^* with loading $\alpha_{m^*,m^*} = 1$ and $\alpha_{m^*,j\neq m^*} = 0$

$$\boldsymbol{\alpha}_{j,m}^{(k+1)} = \frac{\sum_i(\hat{Y}_i^{j(k)} - \boldsymbol{X}_i^{j\prime}\boldsymbol{\beta}^{j(k+1)} - \delta^{j(k+1)})(\hat{Y}_i^{m^*(k)} - (\boldsymbol{X}_i^{m^*})'\boldsymbol{\beta}^{(k+1)} - \delta^{m^*(k+1)})}{\sum_i(\hat{Y}_i^{m^*(k)} - (\boldsymbol{X}_i^{m^*})'\boldsymbol{\beta}^{j(k+1)} - \delta^{m^*(k+1)})^2}.$$

- Update $\theta_i^{(k+1)}$, in this step, we use the closed form solution to update the individual level latent factors, which is more robust than the method proposed in Chen, Fernández-Val, and Weidner (2021).

$$\theta_i^{(k+1)} = (\hat{Y}_i^{j(k)} - \boldsymbol{X}_i^{j\prime}\boldsymbol{\beta}^{j(k+1)} - \delta^{j(k+1)})'\boldsymbol{\alpha}^{j(k+1)}(\boldsymbol{\alpha}^{j(k+1)}\boldsymbol{\alpha}'^{j(k+1)})^{-1}.$$

(4) Iterate until convergence.

[11] Wang's analysis assumes no $\boldsymbol{X}_i^j$ and identifies and develops a consistent estimator of $\boldsymbol{\theta}_i^j$, as well as of factor loadings $\boldsymbol{\alpha}^j$. It is trivial to apply his analysis when components of $\boldsymbol{\theta}_i^j$ are known, i.e., the $\boldsymbol{X}_i^j$.

K Fit of Estimated Models to Sample Data

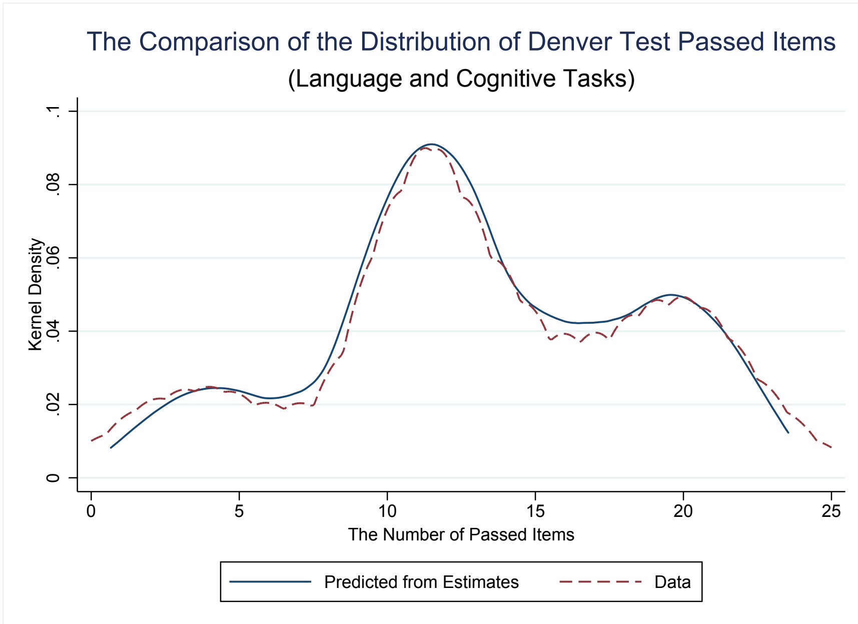


Figure K.1: Model Fit for Language and Cognitive Tasks

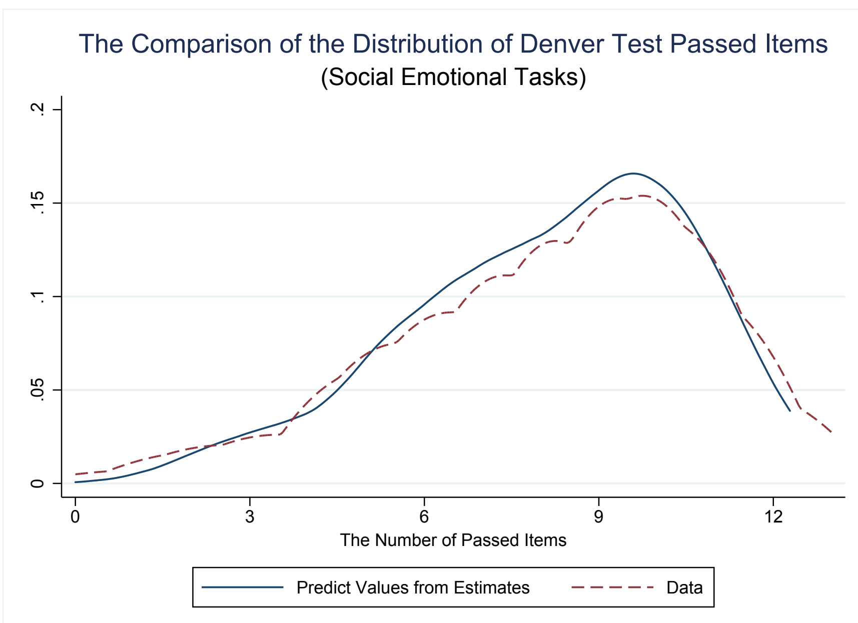


Figure K.2: Model Fit for Social-Emotional Tasks

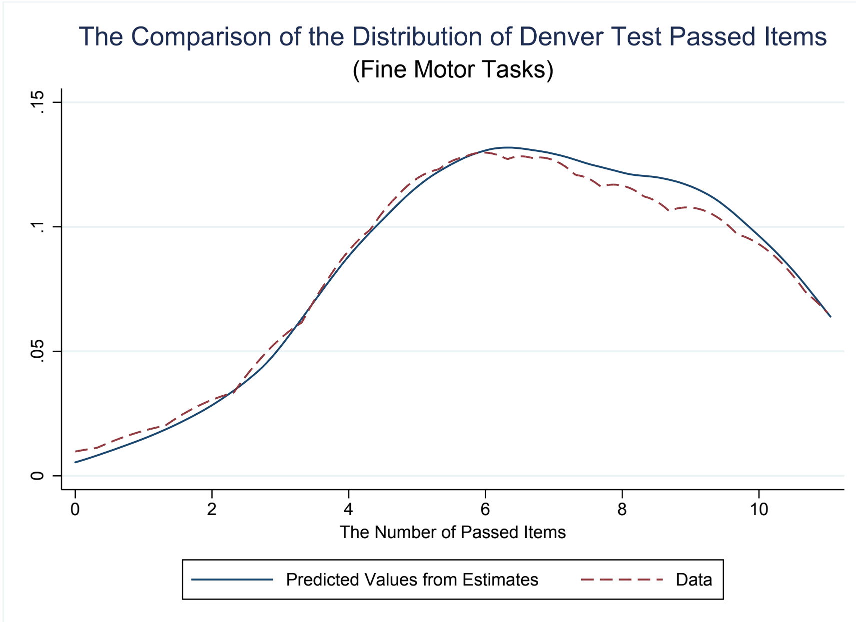


Figure K.3: Model Fit for Fine Motor Tasks

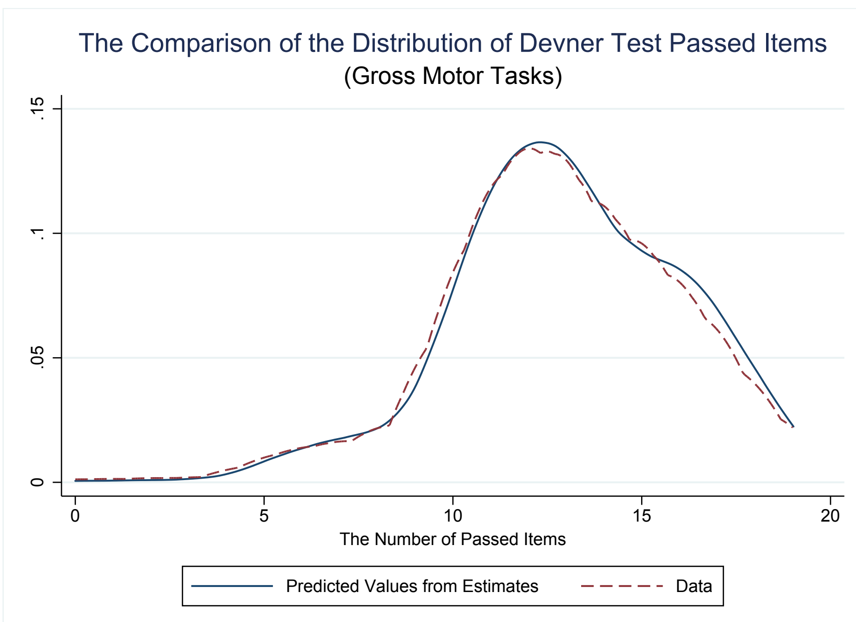
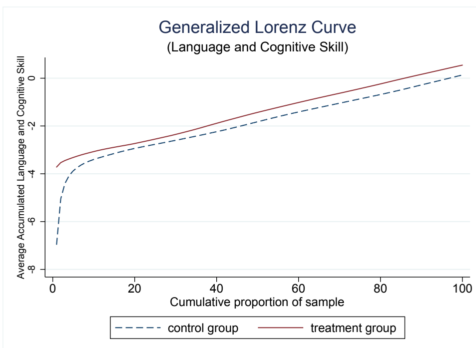


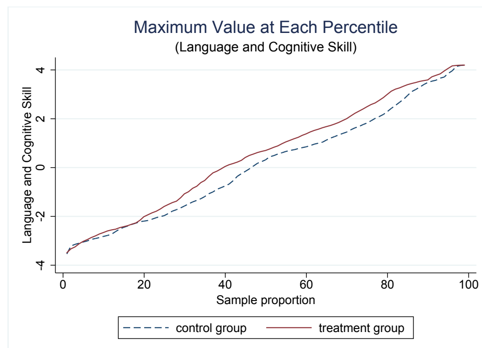
Figure K.4: Model Fit for Gross Motor Tasks

L Stochastic Dominance

We test for stochastic dominance of the estimated skill curves. Figure L.1a, the generalized Lorenz curve, shows the average cumulative values at each cumulative proportion observation. At each cumulative proportion, the treated children have higher language skills. Similarly, Figure L.1b gives the maximum language and cognitive skill values at each percentile. It is clear that the treated group has larger language skill values at each percentile. Figures L.2–L.4 show the same measures for social-emotional, fine motor and gross motor skills, respectively. We can find similar patterns for social-emotional and fine motor skills but not for gross motor skill.

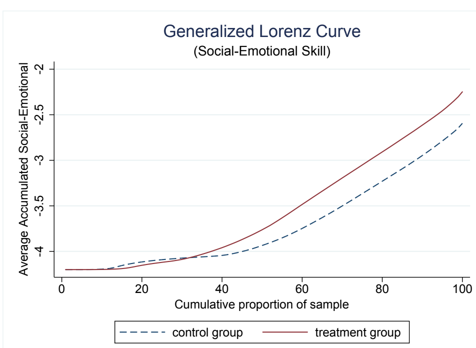


(a)

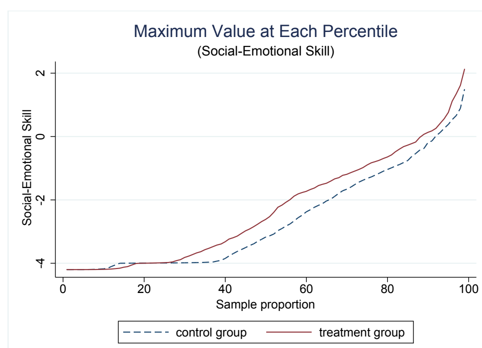


(b)

Figure L.1: Language and Cognitive Skills Stochastic Dominance Curves

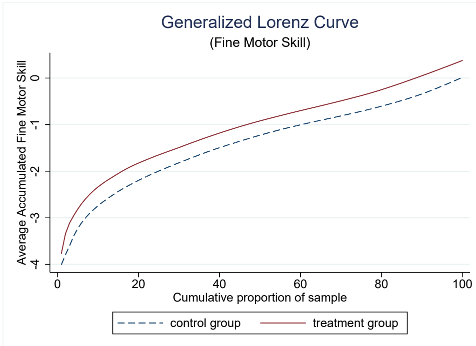


(a)

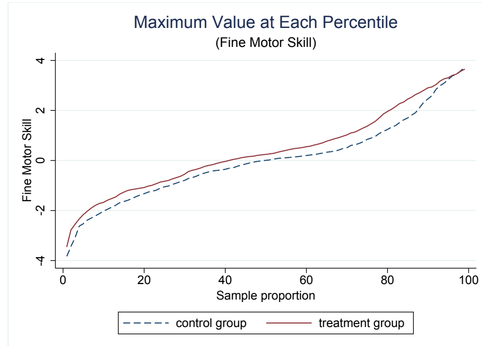


(b)

Figure L.2: Social-Emotional Skills Stochastic Dominance Curves

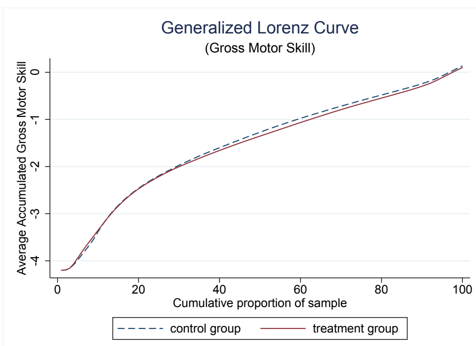


(a)

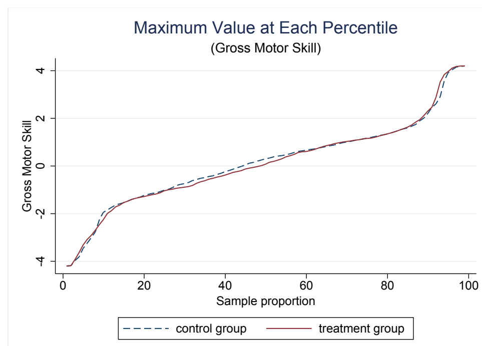


(b)

Figure L.3: Fine Motor Skills Stochastic Dominance Curves



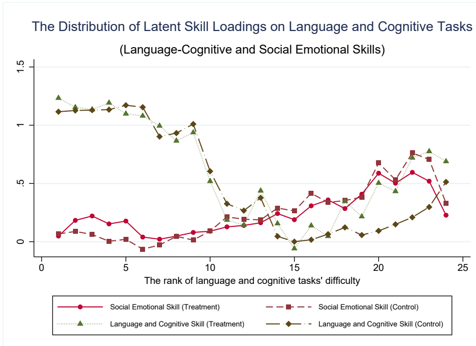
(a)



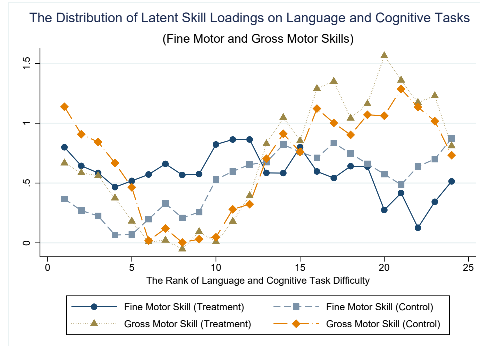
(b)

Figure L.4: Gross Motor Skills Stochastic Dominance Curves

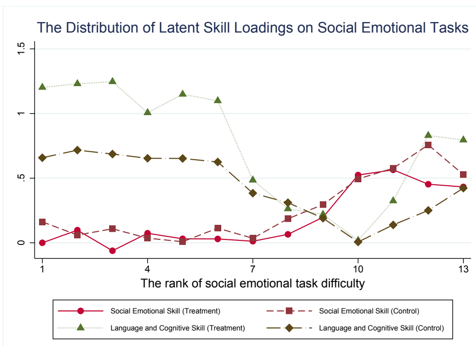
We plot the point estimates of loadings for each task in Figure L.5. We find that the scale of language and cognitive skill loadings on language and cognitive tasks is much higher than the loadings of social-emotional skill on language and cognitive tasks. This finding also holds with other types of tasks. This means that improving language and cognitive skill could boost the child's overall Denver test performance given the same level social-emotional and language skills.



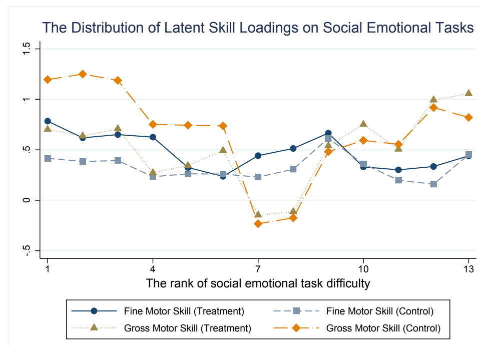
(a)



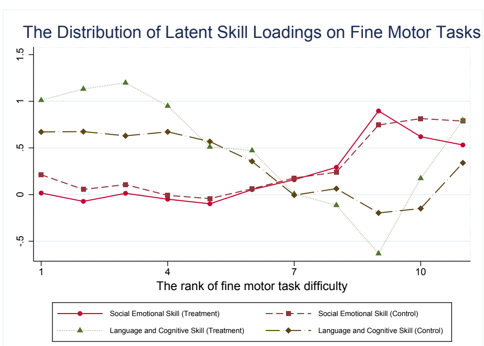
(b)



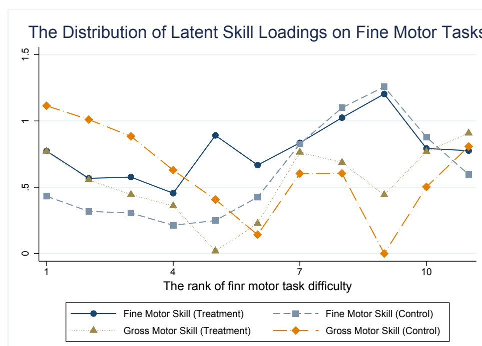
(c)



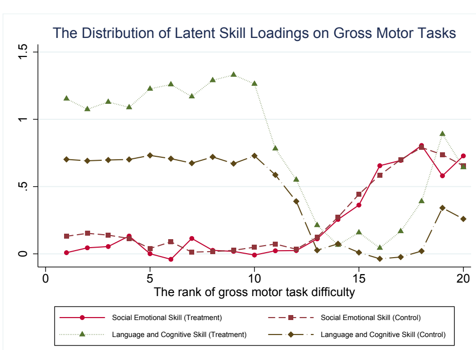
(d)



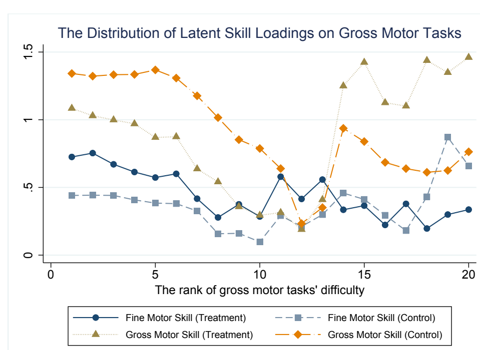
(e)



(f)

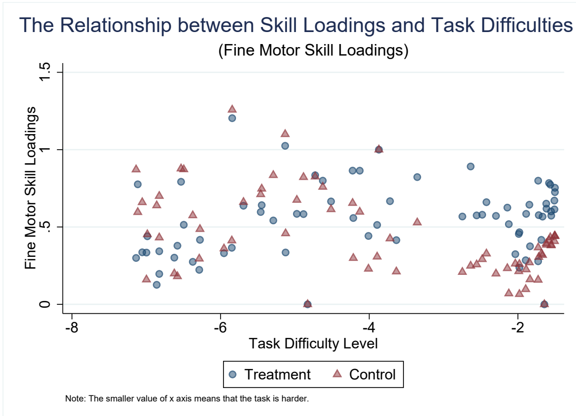


(g)

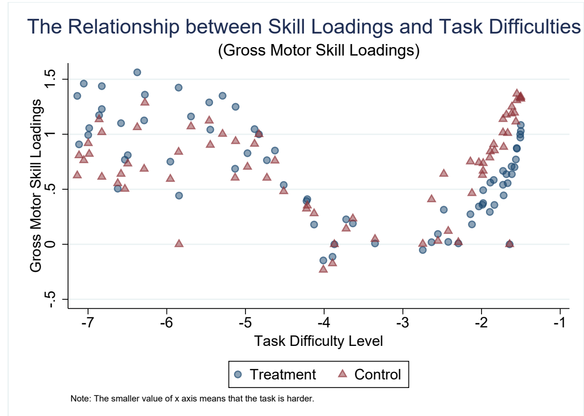


(h)

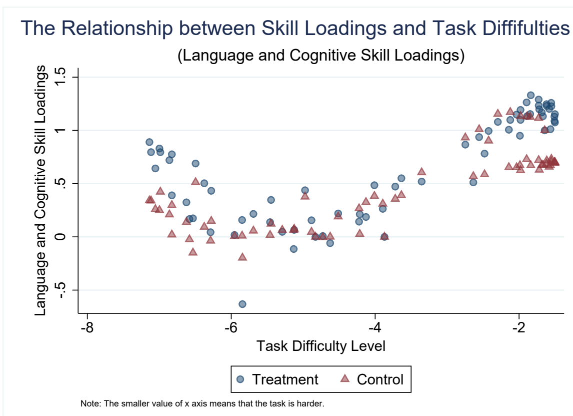
Figure L.5: The Distribution of Latent Skill Loadings



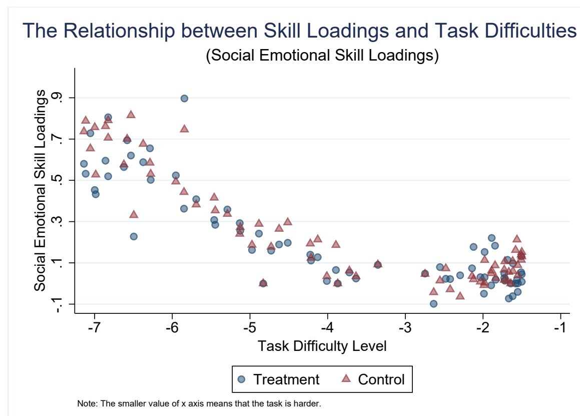
(a)



(b)



(c)

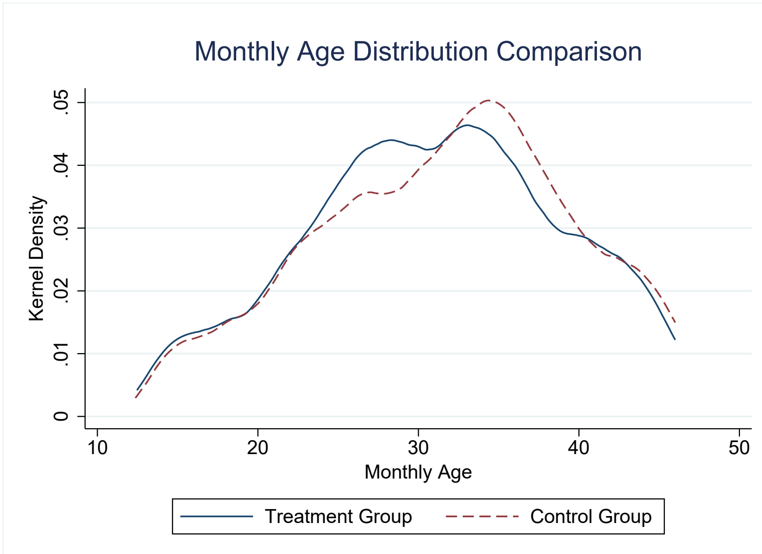


(d)

Figure L.6: The Relationship between Latent Skill Loadings and Task Difficulties

M The Monthly Age Distribution Comparison

Figure M.1: The Monthly Age Distribution Comparison



Note: The p-value of the Kruskal-Wallis test is 0.18.

References

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 90(3), 414–427.

Chen, M., I. Fernández-Val, and M. Weidner (2021): "Nonlinear Factor Models for Network and Panel Data," *Journal of Econometrics*, 220(2), 296–324.

Maasoumi, E., and L. Wang (2019): "The Gender Gap between Earnings Distributions," *Journal of Political Economy*, 127(5), 2438–2504.

Palmer, F. H. (1971): *Concept training curriculum for children ages two to five*. State University of New York at Stony Brook, Stony Brook, NY.

Tsiatis, A. (2006): *Semiparametric Theory and Missing Data*. New York: Springer.

Uzgiris, I. C., and J. Hunt (1975): "Assessment in Infancy: Ordinal Scales of Psychological Development.," *University of Illinois Press*.

Wang, F. (2020): "Maximum likelihood estimation and inference for high dimensional generalized factor models with application to factor-augmented regressions," *Journal of Econometrics*.