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

Jin Zhou[1], James Heckman[1], Bei Liu[2] and Mai Lu[2]

[1]Center for the Economics of Human Development, University of Chicago

[2]China Development Research Foundation

May 1, 2023

*CEHD acknowledges support from the Institute for New Economic Thinking, the Eunice Kennedy Shriver National Institute of Child Health, and Human Development of the National Institutes of Health under award number R37HD065072 and an anonymous donor. 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.

Abstract

This paper estimates the causal impacts on child skills and the mechanisms producing these impacts using data from a randomized control experiment. We study a widely emulated early-childhood home visiting program and show the feasibility of replicating it at scale. We go beyond reporting treatment effects as unweighted item scores and assess item difficulties. To interpret treatment effects, we estimate individual-level latent skills and compare treatments and controls. The program substantially improves multiple skills. We decompose the source of treatment effects and find that enhancements in latent skills explain most of the conventional treatment effects for language and cognition.

JEL Codes: J13, Z18

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

Jin Zhou
Center for the Economics
of Human Development
University of Chicago
1126 East 59th Street
Chicago, IL 60637
Email: jinzhou@uchicago.edu

James J. Heckman
Center for the Economics
of Human Development
University of Chicago
1126 East 59th Street
Chicago, IL 60637
Email: jjh@uchicago.edu

Bei Liu
China Development Research Foundation
Floor 15, Tower A,
Imperial International Center,
No .136 Andingmen Wai Avenue,
Dongcheng District, Beijing
Phone: 86-10-64255855
Email: liubei@cdrf.org.cn

Mai Lu
China Development Research Foundation
Floor 15, Tower A,
Imperial International Center,
No .136 Andingmen Wai Avenue,
Dongcheng District, Beijing
Phone: 86-10-64255855
Email: lumai@cdrf.org.cn

1 Introduction

A growing body of research establishes the effectiveness of home visiting programs targeted to the early years for developing the skills of disadvantaged children. In general, small-scale home visiting programs have been shown to be effective.[1] 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.

This paper studies a replication of the Jamaica Reach Up and Learn program, established over 30 years ago. It is a successful, widely-emulated home visiting program. Visitors have levels of education comparable to those of the caregivers visited. These features facilitate its scalability.[2] Its low cost and flexible format make it an appealing program for less-developed countries.[3]

This paper studies impacts on child skills and parental engagement at midline and endline of a large-scale replication of the original Jamaica program in a poor region of Western China (1500+ participants compared to the 100+ participants in the original Jamaica study). While the curricula are identical, the intake is not. Jamaica targeted stunted children. Its Chinese counterpart enrolls all children in a poor region of the country, save the most biologically compromised. 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 lessons learned from Jamaica into the China replication. The program

[1] See, e.g., Grantham-McGregor and Smith (2016); HomVEE (2020); Howard and Brooks-Gunn (2009).

[2] See List (2022) for a study of taking programs to scale.

[3] See Attanasio et al. (2022); Grantham-McGregor and Smith (2016).

improves home environments. It has a strong impact on language and cognitive skills, fine motor skills, and social-emotional skills, but impacts are not uniform across baseline skill distributions. Positive impacts on skills are strongest for children with absent mothers.

Other research (Heckman and Zhou, 2022a,b,c) uses weekly data on the growth of skills for treatment group children to understand the dynamics of skill formation. They study alternative measures of skill and their relationship to the conventional measures of child development analyzed in this paper. That research is distinct from our analysis of program treatment effects in this paper.

We depart from conventional practice and adjust tests for mastery of tasks for their difficulty across the multiple items used to assess skills. We thereby account for the graduated item difficulty built into the tests we use. We avoid the unjustified but widely followed approach in the literature that reports unweighted counts of performances on tasks that vary in difficulty. Our adjustments produce more plausible estimated treatment effects. Following Heckman et al. (2013), 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.

We estimate *each child's* latent skills instead of just distributions of latent skills as is customary. We decompose treatment effects in order to place program impacts on the population distributions of latent skills.

This paper proceeds as follows: Section 2 describes the program. It is a scaled version of an application of the original curriculum of the Jamaica program. Section 3 presents an array of conventional experimental treatment effects and documents heterogeneity in program impacts. We document beneficial effects of the program on home environments. We augment these estimates with multivariate

factor models to construct individual-level latent skills and determine the impact of treatment on the skills that generate responses to item scores. Visitors who foster positive interactions with caregivers and with children are more effective in promoting cognitive and language skills. Section 4 examines the sources of the estimated treatment effects by examining the extent to which the program affects endowments and the maps between endowments and test scores—a measure of the efficiency of agents in utilizing given stocks of skill. 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

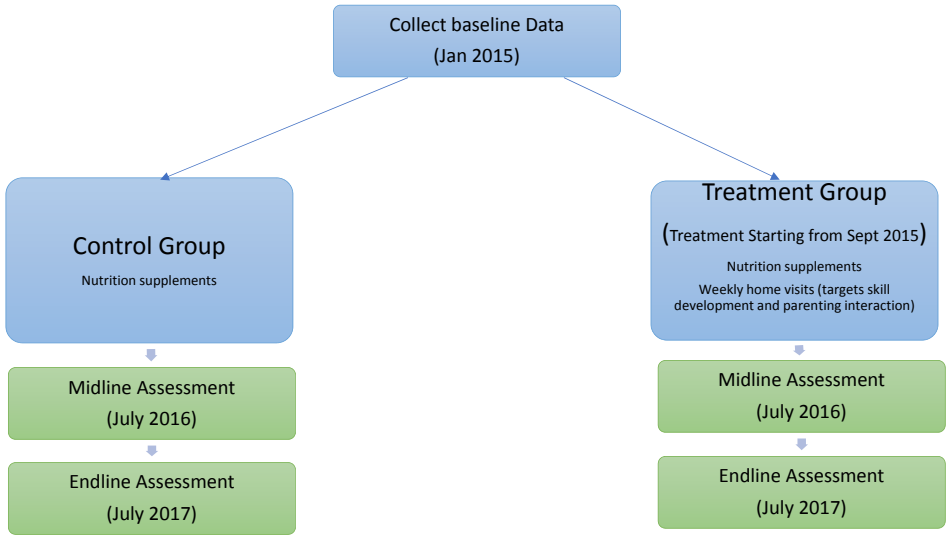
The China Rural Education and Child Health (China REACH) project was launched in 2015 in response to a growing focus on, and call for, evidence-based pilot-to-policy analyses by China's State Council. It is a large-scale randomized control trial (RCT) designed to evaluate the impacts of a low-cost home visit delivery model for disadvantaged families. It is based on the curriculum of a successful Jamaican pilot.[4] 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 in 111 administrative villages. For analytical convenience, two adjacent villages[5] are combined. It is 85%

[4]See Attanasio et al. (2020); Gertler et al. (2022, 2014); Grantham-McGregor and Smith (2016).
[5]Chenghao and Wujiao.

mountainous with a population of 132,000, of whom 114,600 have rural hukou.[6] The program was launched in January 2015 and home visits started in September 2015 (see Figure 1). For details on program implementation, see Appendix A.

Figure 1: 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 child development. Appendix B documents the home visiting protocols used.

[6]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.

Local implementation of the China REACH project is conducted by a county project coordinator, assisted by 24 township supervisors and 91 home visitors.[7] 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 have three years more education than that of the visitors, whose level of education is, on average, at the level of the mothers visited.

The supervisors support and manage the home visitors. They make sure that the home visitors prepare for weekly visits, review the content of past visits, plan activities 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 at least 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.[8] 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 program, the content of each weekly visit, and the assessment instruments used each

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

[8] 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).

week. The curriculum includes more than 200 tasks related to language and cognitive skill development, 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 (2022) shows that this design is optimal for minimizing the mean-squared error of estimates of average treatment effects. We use the experimental design to guarantee exogeneity of regressors and identify the parameters of the underlying structures generating the estimates.

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.[9] In our sample, we have 110 villages. We have 110×110 metrics to measure the distance between each pair of villages. The second step generates 55 pairs and minimizes the sum of Mahalanobis distance of all pairs. To minimize the Mahalanobis metric, we sort the villages by Mahalanobis scores and pair the closest ones using the nonparametric belief propagation (nbp) matching method.[10] The nbp matching method conducts the pairs to minimize the sum of Mahalanobis distance of 55 pairs. Bai (2022) shows that us-

[9]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 (see Appendix Figure C.1); (2) the language skill score 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.

[10]Lu et al. (2011) show that the nbp matching method is optimal and not greedy.

ing Mahalanobis matrix has better performance than other metrics in generating smaller mean-squared errors for average treatment effects.

The third step randomly selects one village within each pair into the treatment group and the other paired village into the control group.[11] Figure A.2 displays the location of the paired villages in Huachi county. The design closely matches the characteristics of the villages in the pairs.[12] Village-level treatment effects include within-village spillovers. Villages are used only once, as treatments or controls. We report treatment effects from a paired matching design, although this is not the main focus of this paper.

3 Estimated Treatment Effects

China REACH 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 gives a detailed assessment of child development.[13,14,15]

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

[12] Appendix C documents baseline comparisons.

[13] 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).

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

[15] 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
Language	Vocalization, gestures, and speaking coherent words.
Fine Motor	The skill of finger movements, such as grasping, releasing and stitching, drawing, and writing.
Social-Emotional	Express and control emotions and communicate in a developmentally appropriate way.
Gross Motor	A wide range of body muscle movements, such as walking, running, throwing, and kicking.

This section reports conventional estimates of the 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 et al., 2008).

Using proportions of items correctly answered as outcomes—the 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 address this problem using a nonlinear measurement model that accounts for item difficulty[16] 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 Estimating Average Treatment Effects

We report the treatment effects for a paired matching design, following Bai et al. (2021) and Bai (2022). Our notation is as follows: The universe of villages is

[16]See, e.g., van der Linden (2016).

$\{1, \ldots, V\}$. 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 $\bar{\boldsymbol{Z}}(\boldsymbol{v})$. Proximity is calibrated by a Mahalanobis metric:

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

where Σ 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.

$D_v = 1$ 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 regression 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 randomized 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., $\varepsilon^m_{iv} \perp\!\!\!\perp \varepsilon^m_{kv'}$ for $\forall i \in v$ and $\forall k \in v', v \neq v'$. Residual plots displayed in Appendix E validate the assumption. 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 $v \neq v'$.[17]

Define the full array of right-hand side variables in Equation (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 biased when $\hat{\Omega}_v$ is estimated using the OLS residuals $\hat{\epsilon}_v$: $E(\hat{\boldsymbol{\epsilon}}_v \hat{\boldsymbol{\epsilon}}'_v)$.[18] The bias depends on the form of $\boldsymbol{\Omega_v}$. Cameron et al. (2008) discuss this problem and show that the wild cluster bootstrap performs well in making cluster-robust inferences. Details of the wild bootstrap procedures used are presented in Appendix F.[19]

In our sample, over 98% of eligible children in the treated villages receive home visits. Still, about 15% of children from both treatment and control groups miss the annual child development assessment. In an effort to obtain consistent estimates of population average treatment effects, we use inverse probability weighting (Tsiatis, 2006).[20,21] In Tables 2–4, we show estimates with IPW and without IPW. In estimating our latent factor model, we also weight the observations.

Table 2 presents the treatment effects for each skill category using standardized

[17] $\boldsymbol{X}_v$ 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.

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

[19] Because we have 55 clusters, recent concerns about the wild bootstrap do not apply. See Canay et al. (2021).

[20] 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.

[21] 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.

outcome measures.[22,23] 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 were 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.[24]

The first row in Table 2 shows that children in the treatment group are, on average, more likely to have higher language and cognitive skills.[25] The first row shows 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 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. In columns (4) and (5), we restrict the sample to children who were less than 2 years old at enrollment. This helps to generate a more age-balanced sample between treatment and control groups. In Appendix Figure H.1, we show that the monthly age distributions are comparable between treatment and control groups.

The intervention significantly improves social-emotional skills at midline and fine motor skills at the end of the intervention but produces no significant improve-

[22]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.

[23]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 treatment and control groups.

[24]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.

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

ment in gross motor skills. This finding is consistent with the design of the curriculum, which focuses primarily on language and cognitive skill development.[26,27]

Tables 3–4 display treatment effects by gender. An interesting finding, consistent with recurrent findings in the literature (Elango et al., 2016), is that the intervention improves boys' language and cognitive skills much more than that of girls. At midline, the treatment effect sizes are 0.4 for girls and 0.9 for boys. At the end of the intervention, the effect sizes are 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 social-emotional skills.[28]

[26] Heckman and Zhou (2022a) document the intervention curriculum.

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

[28] This result is also found in the evaluation of the Perry Preschool program (García et al., 2018) and the Abecedarian preschool program (García et al., 2018).

Table 2: Treatment Effects on Standardized Denver Scores

	(1) All	(2) All	(3) All	(4) Children $\leq$ 2 Yrs at Enrollment	(5) 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) All	(2) All	(3) All	(4) Children $\leq$ 2 Yrs at Enrollment	(5) 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) All	(2) All	(3) All	(4) Children $\leq$ 2 Yrs at Enrollment	(5) 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.

Our estimates are robust when we use the matching estimation method instead of IPW-weighted OLS. These results are reported in Appendix I.

The program was designed to improve the home lives of treated children. Data are collected for both treatment and control groups on home environments as assessed by supervisors. Table 5 reports the treatment effects on HOME environment scores. The intervention significantly improves the composite HOME score and parental involvement at home. Data are also collected *for the treatment group* on the quality of interactions between visitors and caregivers and between visitors and children. Appendix J discusses the measures of interaction at our disposal and how we combine measures using an empirical Bayes procedure to form our interaction variables.[29] Tables J.3–J.6 report estimates of the impacts on child skills of measured interactions between the home visitor and caregivers, and the home visitor and children, as well as the impact of home visitor's teaching ability.[30] The only strong pattern that emerges is that good caregiver–home visitor interactions promote language and cognitive skills for children.[31]

[29] Table J.1 in Appendix J lists the data collected.

[30] 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.

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

Table 5: Treatment Effects on Home Environment Scores

	(1)	(2)	(3)	(4)	(5)
	All	All	All	Children $\leq$ 2 Yrs at Enrollment	Children $\leq$ 2 Yrs at Enrollment
Home Total	0.641**	0.892***	0.868***	0.705**	0.72**
	[0.127, 1.163]	[0.338, 1.459]	[0.309, 1.409]	[0.142, 1.276]	[0.159, 1.269]
Home Involvement	0.184***	0.217***	0.241***	0.169**	0.201***
	[0.063, 0.309]	[0.093, 0.342]	[0.109, 0.367]	[0.048, 0.288]	[0.073, 0.327]
Home Variety	0.093	0.118	0.114	0.091	0.093
	[-0.029, 0.215]	[-0.008, 0.248]	[-0.025, 0.253]	[-0.035, 0.213]	[-0.037, 0.224]
Home Responsivity	-0.001	0.079	0.066	0.056	0.048
	[-0.207, 0.213]	[-0.145, 0.304]	[-0.169, 0.300]	[-0.175, 0.282]	[-0.192, 0.289]
Home Acceptance	0.069	0.083	0.059	0.067	0.044
	[-0.018, 0.152]	[-0.012, 0.178]	[-0.041, 0.157]	[-0.037, 0.170]	[-0.064, 0.150]
Home Organization	0.116	0.149	0.095	0.116	0.069
	[-0.031, 0.263]	[-0.003, 0.296]	[-0.059, 0.242]	[-0.038, 0.270]	[-0.077, 0.223]
Home Learning Materials	0.18	0.245	0.291*	0.205	0.262*
	[-0.064, 0.422]	[-0.019, 0.504]	[0.047, 0.533]	[-0.063, 0.472]	[0.007, 0.512]
Pre-Treatment Covariates	No	No	Yes	No	Yes
IPW	No	Yes	Yes	Yes	Yes

Notes: 1. The 90% confidence intervals in brackets are constructed using the wild bootstrap clustered at the village level.
2. 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.
3. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

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

The previous analysis shows that the treatment boosts outcomes on unweighted item aggregates. Aggregates so formed, while traditional, are problematic unless the difficulty is the same across all 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.[32] We follow standard methods in psychometrics and introduce and estimate difficulty parameters across items.[33] We also estimate individual-level latent skills. We use our estimates to determine the impact of treatment on the skills that generate item scores. Following Heckman et al. (2013), 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, language, 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 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

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

[33] van der Linden (2016).

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

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

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

where $\boldsymbol{\alpha}^{j_k,d}$ 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 $\boldsymbol{\beta}^{j_k,d}$ and $\boldsymbol{\alpha}^{j_k,d}$ can depend on treatment, the skills modeled, and even the item studied, where items are common across people. In estimation, we impose $\boldsymbol{\beta}^{j_k,d} = \boldsymbol{\beta}^{j'_k,d} = \boldsymbol{\beta}^{k,d}$, $\forall j_k$ and j'_k; i.e., coefficients are common across all items. Specification (3) generalizes the scalar Rasch model to allow vector skills.[34]

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 $\boldsymbol{\alpha}^{j_k,d}$. The expression $(\boldsymbol{\theta}_i^d)'\boldsymbol{\alpha}^{j_k,d}$ 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 $\boldsymbol{\theta}_i^d$ 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

[34] See van der Linden (2016) for a discussion of the Rasch model.

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. The analysis of Wang (2020) shows 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.[35] These conditions apply in our sample with large numbers of test items per person ($\geq$ 70 for each skill) and large numbers of observations.

Factor models require normalizations if one seeks to isolate $\boldsymbol{\theta}^{\boldsymbol{d}}$ from $\boldsymbol{\alpha}^{j_k,d}$. Since $(\boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{d}})'\boldsymbol{\alpha}^{j_k,d} = (\boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{d}})'\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{\alpha}^{j_k,d}$, 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, $(\boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{d}})'\boldsymbol{\alpha}^{j_k,d}$. However, it is also interesting to break out the impact of the intervention on each component.

We do so using a normalization suggested by Anderson and Rubin (1956) and identify both the vectors $\boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{d}}$ and $\boldsymbol{\alpha}^{j_k,d}$.[36] This enables us to examine the impacts of the intervention on endowments and the impacts of the intervention on the efficiency of agents in using skills. We report estimates for $\boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{d}}$ and $\boldsymbol{\alpha}^{j_k,d}$ separately and also as a bundle of effective skills $(\boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{d}})'\boldsymbol{\alpha}^{j_k,d}$.

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-

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

[36] We provide the details of Anderson and Rubin's (1956) normalization method in Appendix K.

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 generalizes what Cunha et al. (2010) call the "dedicated factor case" to apply to only the first four items of each measurement. We only require that 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 $\boldsymbol{I_{K,K}}$ identity matrix. For the first $K = 4$ items of the measurements, we assume that they load on only one skill.[37] 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 $\boldsymbol{\alpha}'_{N_J \times K}$:

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

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 $\ell \in \{1, \ldots, 4\}$. Table 6 reports this test. The assumption of a dedicated factor model fails in our sample.

We report sensitivity analyses of our estimates using a variety of plausible normalizations in Appendix L. We find that the estimates of $\boldsymbol{\alpha}^{j_k,d}$ reported in the text

[37] 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.

Table 6: Test of Hypothesis $\alpha^{j_k,\ell,d} = 0$ for $j_k \geq 5$ except for one $\ell \in \{1,\ldots,4\}$

	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	1186.793	0.000	2108.501	0.000
Gross Motor	1570.322	0.000	1969.099	0.000

are stable under a variety of different normalizations.[38] Our results are quantitatively robust. We use the estimation procedure proposed by Chen et al. (2021) to estimate panel probit models with multiple latent skill factors.[39]

3.2.2 Estimates

Table 7 presents estimates of $\boldsymbol{\beta}^{k,d}$. There are no statistically significant differences between the treatment and control groups, although the point estimates for 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.[40] We fit the data as well when we investigate the other types of tasks.[41] We find qualitatively similar results when we use a richer set of covariates. See Appendix Table O.1.

[38] In Appendix L, 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.

[39] Details regarding the method are presented in Appendix M. The asymptotic justification for this approach for estimating individual-specific factors and population factor loadings is based on Wang (2020).

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

[41] See Appendix N.

Table 7: Estimated Coefficients for 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 χ^2 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.

Figure 2: The Distribution of Denver Test Passed Items

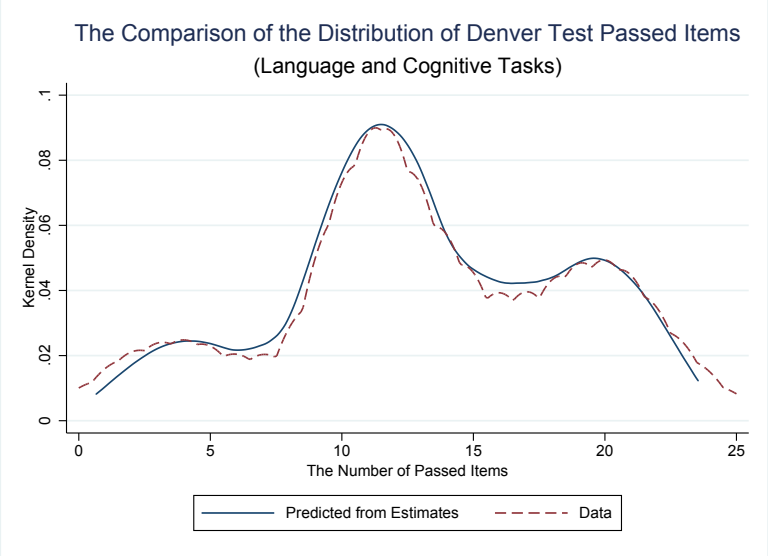
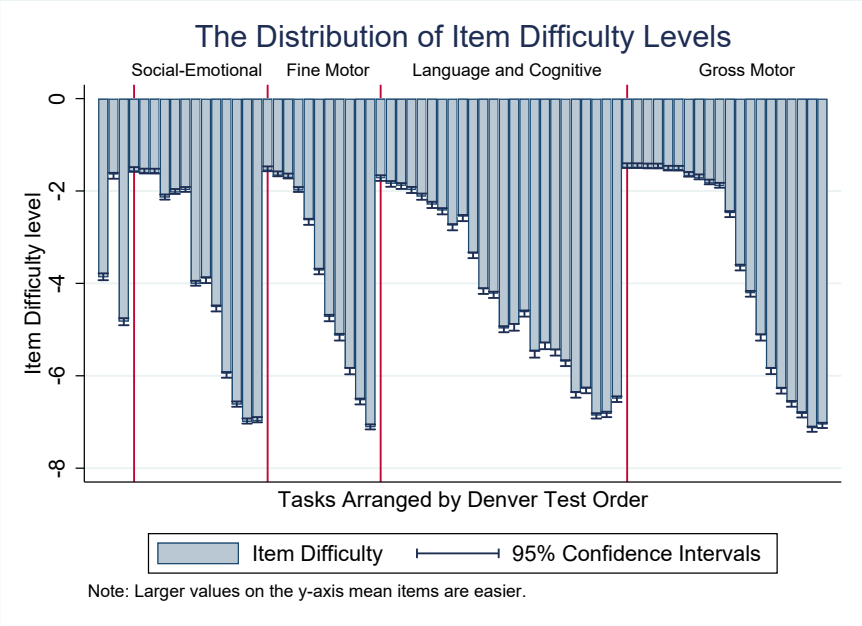


Figure 3 shows the array of estimated difficulty level parameters δ^{jk} for each task item. When the item difficulty level increases, the estimates become more negative. The estimates generally accord with the design of tests to increase the

difficulty level with later items. The estimated difficulty level parameters δ^{jk} 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 can correctly answer easy items but were likely to fail 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.

Figure 3: The Distribution of Denver Task Item Difficulty Levels



An advantage of our approach is that we can estimate individual-level latent skill factors. First, Table 8 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 statistically significantly higher than those in the control group. When we compare treatment effects across different latent skills,

Table 8: 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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: The Correlation between Different Latent Skill Factors

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

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

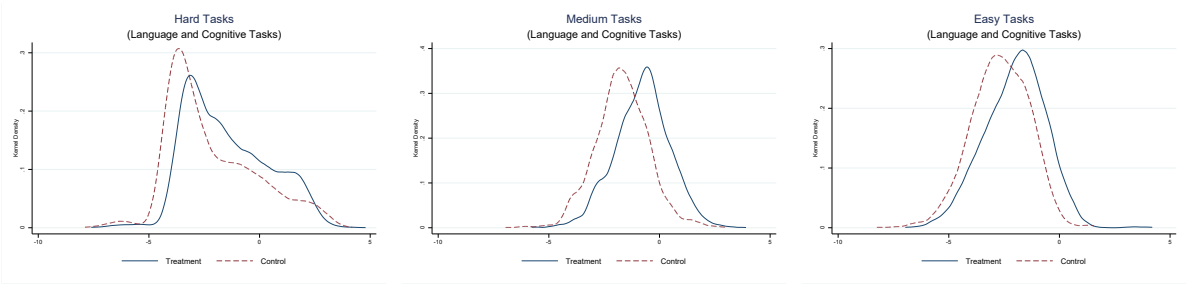
we find that improvements in fine motor and language skills are at the same level but that there are no treatment effects for gross motor skills. Table 9 shows that language and cognitive skills are negatively correlated with gross motor skills and positively correlated with social-emotional and fine motor skills.

Identifying factors and factor loadings is fraught with controversy regarding appropriate normalizations. Figure 4 plots effective skills—the product of estimated skill factor loadings and the latent skill factors $\boldsymbol{\alpha}'\boldsymbol{\theta}$ based on the Denver task difficulty levels for language and cognitive skills.[42] This term does not require any normalization assumptions. On average, the loadings for the treatment group are larger for all tasks whatever their difficulty, but the shift for the loadings of easy tasks is less clear. Figures P.4–P.6 in Appendix P present the comparison of the distribution of $\boldsymbol{\alpha}'\boldsymbol{\theta}$ for treatment and control groups for other skills. The same pattern emerges. Effective skills are increased regardless of any normalization.

[42]Figures P.1 and P.2 present 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 rankings are based on the estimates of the task difficulty level parameters.

Figure 4: Distributions of $\left[(\theta_i^d)'\alpha^{j_k,d}\right]^{\dagger}$ for Language and Cognitive Tasks

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



† Easy tasks are defined as the bottom 33% of all language and cognitive tasks ordered by difficulty level estimates, medium tasks are those that fall between 33% and 66% of all the language and cognitive tasks ordered by difficulty level estimates, and hard tasks are the top 66% of all the language and cognitive tasks ordered by difficulty level estimates.

When we invoke our normalization, for most tasks, we reject the hypothesis that on average factor loadings are the same across treatment and control groups.[43] Table 10 reports tests of equality of the average loadings on different tasks for the different skills. Except for gross motor skills, we reject the hypothesis. The loadings on latent language and cognitive skills are large, but the loadings for social-emotional skills are smaller, suggesting that, on average, the program reduces the effectiveness of such skills.

We also test equality of the vector $\alpha^{j_k,\ell,d=1} = \alpha^{j_k,\ell,d=0}$. In Appendix P, Tables P.1–P.2 present such tests. While we cannot reject equality for social emotional loadings jointly, we can reject equality for the other types of skill loadings.

[43] Tables P.1–P.2 provide item-by-item tests. Social-emotional item loadings are not precisely estimated.

Table 10: Estimated Skill Loadings on Denver Test Tasks ($\boldsymbol{\alpha}^{j_k,\boldsymbol{d}}$) Latent Skills

Control			Treatment			p-value
Skill Loadings	Mean	S. D.	Skill Loadings	Mean	S.D.	test of equality of mean loadings
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 $\boldsymbol{\alpha}^{j_k,\mathbf{0}}$ and $\boldsymbol{\alpha}^{j_k,\mathbf{1}}$, respectively, across items.
2. p-values are for the null of equality of treatment and control summary measures.

3.2.3 Comparisons with a Model without Task Difficulty Parameters

To show the impact of introducing task difficulty parameters into 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 task difficulty across items is less than 0.001.

Second, we compare the treatment effects on the mean of latent skill factors in Table 11 ($E(\boldsymbol{\theta}^1) - E(\boldsymbol{\theta}^0)$). Notice that 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 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 11: Comparing Treatment Effects of θ_i Based on Two Models with and without Difficulty Parameters

	Social-Emotional	Fine Motor	Language	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]

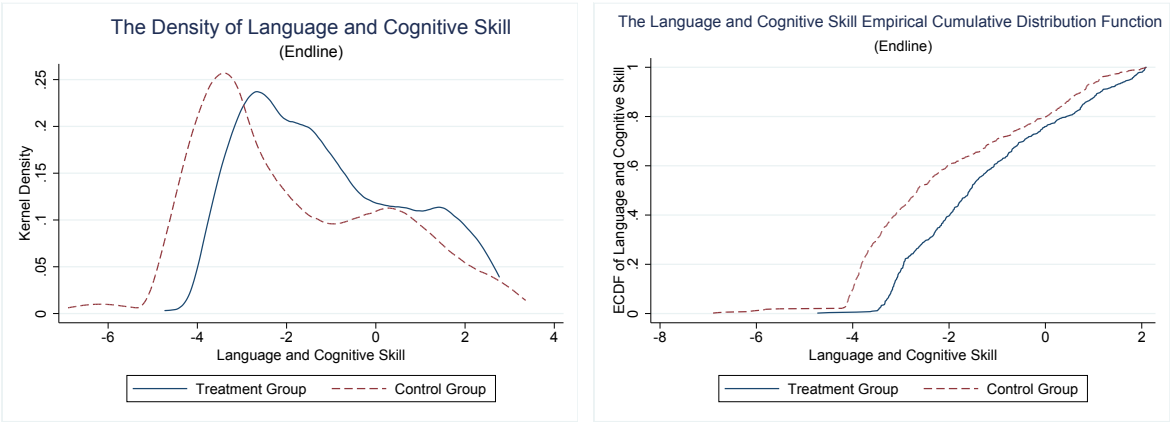
Notes: 1. The 95% confidence intervals in brackets are constructed by wild bootstrap clustered at the village level.
2. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.2.4 Distributions of Latent Skill

We compare the language skill distributions of the treatment and control 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 the same for cumulative distributions. Latent language and cognitive skill distributions are more right-shifted for the treatment group. Differences are more substantial at the bottom and middle of the treatment distribution compared to those at the top.

Figure 5: Language Skills Distribution

(a) (b)



Figures 6a and 7a present the densities of social-emotional and fine motor skills, respectively. For social-emotional skills, skills for the treated are more right-shifted at the top. For fine motor skills, there is a more uniform shift for the treated.

Figure 6: Social-Emotional Skills Distribution

(a) (b)

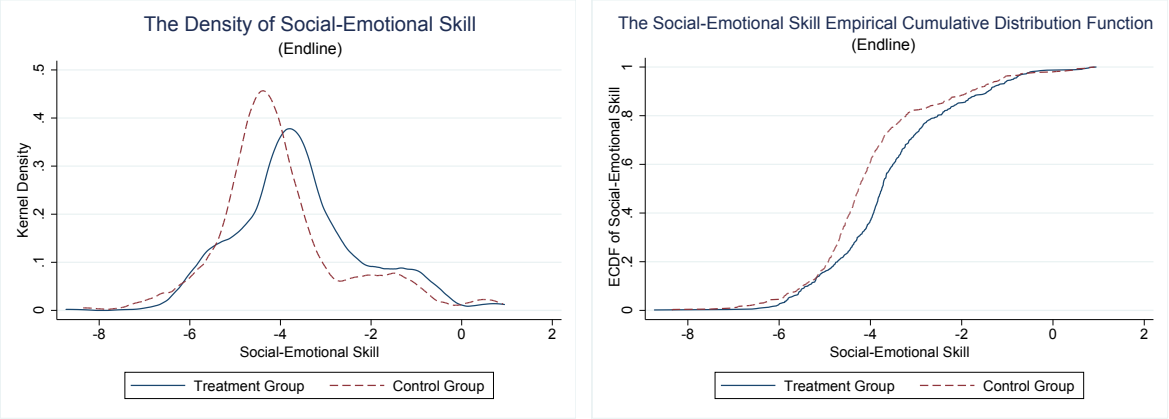
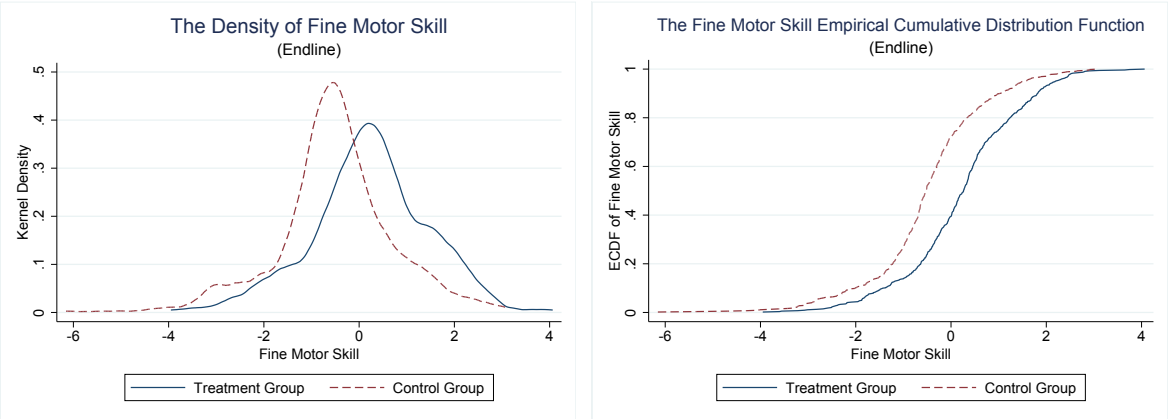


Figure 7: Fine Motor Skills Distribution

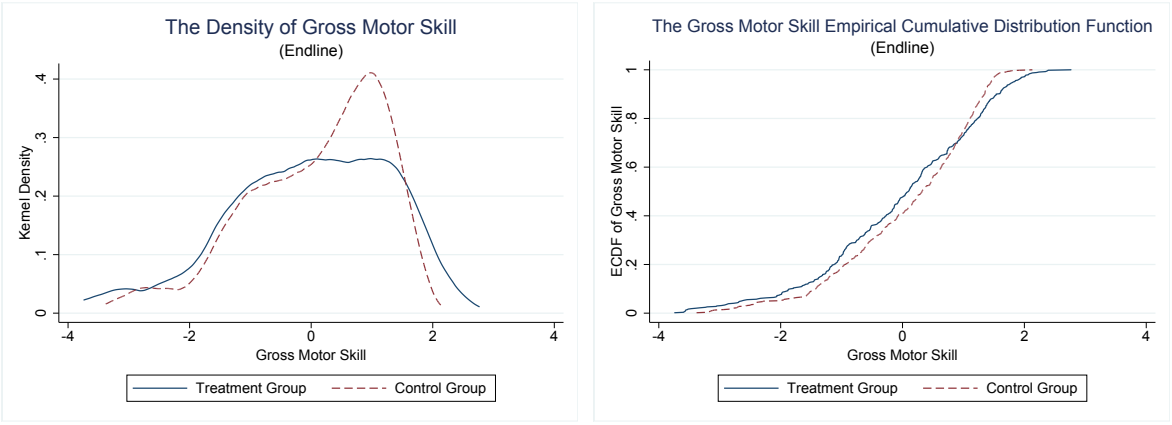
(a) (b)



For gross motor skills, there is little evidence of any treatment effect. The factor distributions are similar between the treatment and control groups. Figures 8a and

Figure 8: Gross Motor Skills Distribution

(a) (b)



8b show that the densities and CDFs of the two gross motor skills distributions are close to each other.

In summary, language and cognitive, social-emotional, and fine motor skills were substantially improved by the program. Assuming a perfect ranking across counterfactual distributions, gains are not uniform across the control distribution for language and cognitive skills. They are roughly 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 Q presents an extensive 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. 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. Before doing so, we assess the performance of our skill estimates in predicting experimental treatment effects.

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 in order to gauge the quality of our estimates. The point estimates of the average treatment effects so obtained are in close agreement (see Table 12).

4.1 The Sources of Our 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 $\boldsymbol{\alpha}^{j_k,d}$ and $\boldsymbol{\beta}^{j_k,d}$. We previously established treatment shifts factor loadings. At issue is whether such shifts explain a quantitatively important portion of estimated treatment effects. To do so, we decompose the item-level treatment effects into two components: the effects from the changes in the mapping from skills to tasks and

Table 12: 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.

the effects of treatment on skills.

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

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

where we assume $\varepsilon_i^{j_k} \sim N(0,1)$. Home visiting treatment effects arise 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 $F^1(\boldsymbol{\theta^1, X})$ and $F^0(\boldsymbol{\theta^0, X})$ as the distributions of $(\boldsymbol{\theta^1, X})$ and $(\boldsymbol{\theta^0, X})$ in the treatment and control populations, respectively. Population treatment effects for item j_k can be decomposed as follows:

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

Notice that Equation (6) holds over a common support for $\boldsymbol{X}$ and when the factors in the treatment and control groups have similar distributions of observable covariates, which are essentially satisfied in our sample.[44] Table 13 reports the decomposition of treatment effects. The main drivers of the treatment effects are increases in latent skills. We previously demonstrated that there is no significant difference in $\boldsymbol{\beta}$ between the treatment and control groups in Table 7. The contribution to the treatment effects from $\boldsymbol{\beta}$ is insignificant, and we ignore it. We decompose the treatment effects in the order suggested in Equation (6). The contribution from experimentally induced changes in $\boldsymbol{\alpha}$ is not precisely estimated, despite the statistically significant shift in the $\boldsymbol{\alpha}$s documented in Table 10. For this reason, we conclude that the dominant effect of treatment is on latent skills. Section S in the appendix shows decompositions conducted in different orders for different sets of family conditioning variables, which lead to similar qualitative and quantitative results.

[44] 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 R, we show the age distribution between treatment and control groups.

Table 13: Source of Treatment Effects (Decompose Observed Covariates First)

	Total Net Treatment Effects	From Observable Covariates	From Skill Loadings	From Latent Skills
Language and Cognitive	1.143	-0.058	0.217	0.984
	(0.185)	(0.190)	(0.192)	(0.188)
		-5%	19%	86%
Social-Emotional	0.239	-0.163	0.049	0.354
	(0.083)	(0.086)	(0.088)	(0.084)
		-68%	20%	148%
Fine Motor	0.317	-0.016	-0.003	0.336
	(0.085)	(0.088)	(0.090)	(0.088)
		-5%	-1%	106%
Gross Motor	0.164	-0.054	0.062	0.156
	(0.100)	(0.106)	(0.109)	(0.103)
		-33%	38%	95%

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 the number 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.

4.2 Treatment Effects on Latent Skills Conditional on Caregiver Status

This section compares treatment effects based on the caregiver's presence status with children. 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 14 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 14 reveals that the largest treatment effects are for vulnerable children for whom mothers are absent (i.e., mother works outside or both parents work outside). Heckman and Zhou (2022c) show that, in most cases, grandmothers with low levels of education are the caregivers when mothers are absent.

Table 14: Treatment Effects on Latent Skills $\boldsymbol{\theta_i}$

Standardized	(1) Non-Left-Behind Children	(2)	(3) Left-Behind Children	(4)
		Mother Works Outside	Father Works Outside	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 15 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 (Gertler et al., 2014; Grantham-McGregor and Smith, 2016). 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 15: Treatment Effects on China REACH and Jamaica Reach Up and Learn

	Panel A: China REACH (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 Home Visiting (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 policies in China are driven by data, rigorous evidence on China REACH has the potential to have a large effect on policy discussions.

Our analysis offers a prototype for measuring latent skills using diverse outcome measures adjusting for the difficulty inherent in different tasks. We estimate child latent skills and how they are affected by the program. We develop a framework for understanding the mechanisms generating treatment effects. We adjust for the difficulty of the various tasks used to assess performance in the program. Such adjustments produce more plausible estimates. We test and reject the "dedicated factor" measurement model widely used in the economics of skill formation. Measured item scores depend on multiple skills.

The intervention improves the quality of home life for children. It significantly boosts children's cognitive and language, fine motor, and social-emotional skills. Program impacts are not uniform across baseline skill levels and are largest for the most vulnerable children. Improvements in latent skills are the dominant component of estimated treatment effects.

References

Anderson, T. W. and H. Rubin (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 5, Berkeley, CA, pp. 111–150. University of California Press.

Attanasio, O., J. Behrman, M. Day, S. Grantham-McGregor, P. Gupta, P. Jervis, P. Makkar, C. Meghir, R. Pal, A. Phimister, and N. Vernekar (2022). Effects of early stimulation and enhanced preschool: Pathways to effective learning. Unpublished Manuscript.

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2020). Estimating the production function for human capital: Results from a randomized controlled trial in Colombia. *American Economic Review 110*(1), 48–85.

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *Conditionally accepted by the American Economic Review*.

Bai, Y., J. P. Romano, and A. M. Shaikh (2021). Inference in experiments with matched pairs. *Journal of the American Statistical Association*.

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 (2021). The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics*, 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, May). 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 R. A. Moffitt (Ed.), *Economics of Means-Tested Transfer Programs in the United States*, Volume 2, Chapter 4, pp. 235–297. Chicago: University of Chicago Press.

García, J. L., J. J. Heckman, D. E. Leaf, and M. J. Prados (2018). Quantifying the life-cycle benefits of a prototypical early childhood program. Forthcoming at the *Journal of Political Economy*, 2020.

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, S. M. Chang, S. Grantham-McGregor, C. Vermeersch, S. Walker, and A. S. Wright (2022). Effect of the Jamaica early childhood stimulation intervention on labor market outcomes at age 31. NBER Working Paper 29292. Under Revision.

Gertler, P., J. J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. M. Chang, and S. 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. and J. Zhou (2022a). Interactions as investments: The microdynamics

and measurement of early childhood learning. Under revision, *Journal of Political Economy*.

Heckman, J. and J. Zhou (2022b, April). Measuring knowledge. Working Paper 29990, National Bureau of Economic Research.

Heckman, J. and J. Zhou (2022c). Nonparametric tests of dynamic complementarity. Unpublised manuscript, University of Chicago.

Heckman, J. J., R. Pinto, and P. A. Savelyev (2013, October). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 2052–2086.

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.

List, J. A. (2022). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. New York: Currency.

Lizzeri, A. and M. Siniscalchi (2008, August). 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. *The American Statistician 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*.