Treatment Effects and the Measurement of Skills in an Influential Home Visiting Program*

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

April 29, 2021

*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 efforts on program design, implementation, and data cleaning support of the China REACH study. Erlfang Tsai and Fuyao Wang provided highly competent research assistance. CDRF thanks Mary Young, Fan Bu, Peng Liu, Lijia Shi, Bojiao Liang, 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.

Corresponding author: Jin Zhou (jinzhou@uchicago.edu)

Abstract

This paper evaluates the causal impacts of an early childhood home visiting program for which treatment is randomly assigned. We estimate multivariate latent skill profiles for individual children and compare treatments and controls. We identify average treatment effects of skills on performance in a variety of tasks. The program substantially improves child language and cognitive, fine motor, and social-emotional skills development. Impacts are especially strong in the most disadvantaged communities. We go beyond reporting treatment effects as unweighted sums of item scores. Instead, we examine how the program affects the latent skills generating item scores and how the program affects the mapping between skills and item scores. We find that enhancements in latent skills explain at least 80% of conventional unweighted treatment effects on language and cognitive tasks. The program enhances some components of the function mapping latent skills into item scores. This can be interpreted as a measure of enhanced productivity in using given bundles of skills to perform tasks. This source explains at most 20% of the average estimated treatment effects.

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 in developing the skills of disadvantaged children. Home visiting programs have previously been shown to be effective (see, e.g., Howard and Brooks-Gunn, 2009; HomVEE, 2020; Grantham-McGregor and Smith, 2016) and 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 some 30 years ago, is a successful prototype of a home visiting program emulated around the world.

This paper studies a close replica of the original Jamaica Reach Up and Learn program, China REACH, which was brought to scale 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 and incorporated their experience into the China version. We show that it has a strong impact on language and cognitive skills, fine motor skills, and social-emotional skills. Impacts are especially strong in the most disadvantaged communities.[1]

In achieving these results, we adjust for task difficulty across the multiple items used to assess skills and thus avoid the unjustified approach widely followed in the literature of reporting unweighted counts of performances on tasks, which vary in difficulty. Doing so produces more plausible estimated treatment effects. We decompose estimated treatment effects into induced improvements in latent skills and improvements in the technology

[1] In Appendix I, at baseline, village-level income per capita and the HOME environment verbal skill and learning material scores are significantly worse for the top ten-performing villages than for the low-performing villages at the 10% level in Table I4.

mapping skills into performance on tasks. Treatment effects mainly arise from boosts in skills. At least 80% of the estimated treatment effects are due to changes in latent inputs with the rest attributable to improvements in the maps between skills and outcomes.

This paper proceeds as follows. Section 2 describes the program and places it in context as a scaled and enhanced version of an original program. Section 3 presents an array of experimental treatment effects. We document heterogeneity in impacts. Section 4 examines the sources of the estimated treatment effects. Following Heckman et al. (2013), we examine whether the program affects the inputs in the functions mapping skills to performance on tasks and whether it shifts the productivity of the map of latent skills to item responses. Section 5 compares the outcomes from the China program with those from the successful parent Jamaican program where followup is through age 30. China REACH is on track to succeed in the long-term improvement of education and labor market outcomes. Section 6 summarizes our findings. Supporting material is reported in a web appendix: http://cehd.uchicago.edu/china-reach_home-visiting_appendix.

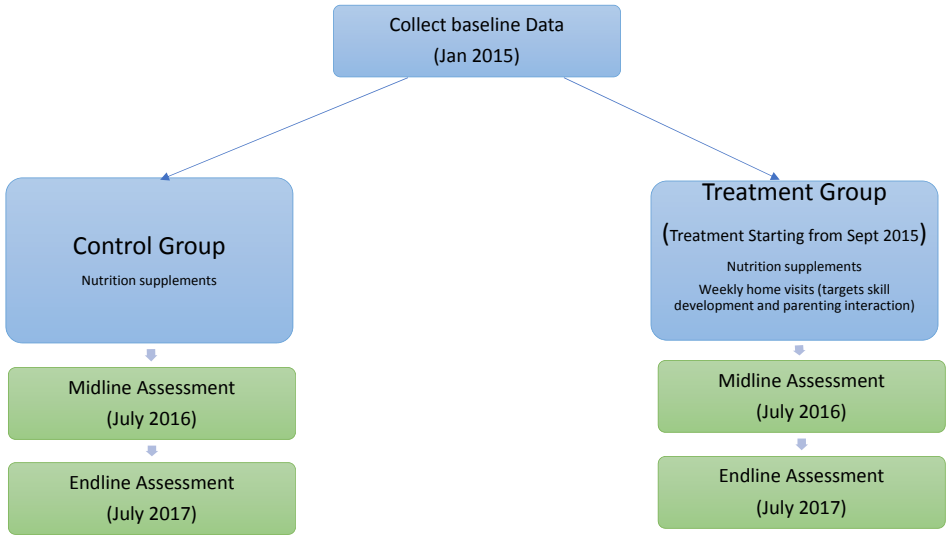
2 China REACH

The ongoing Rural Education and Child Health (China REACH) project was initially launched in 2015 in response to a growing focus on, and call for, evidence-based pilot-to-policy analyses by China's State Council. China REACH is a large scale randomized control trial (RCT) designed to evaluate the impact of a home visit delivery model for disadvantaged families. It is based on the successful Jamaican pilot (e.g., Grantham-McGregor and Smith, 2016; Gertler et al., 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, which is one of the poorest areas in China. The county has 15 townships, including 111 administrative

villages. 85 percent of the county is mountainous. The population is 132,000 people, of whom 114,600 have rural hukou.[2] In Figure 1, we show that the version of the program we study was started in January 2015, and home visits started in September 2015. For details of program implementation, see Appendix A.

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



2.1 The Intervention Implemented

The program trained 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 so large. The program encourages child caregivers to interact with their children in developmentally appropriate ways. Heckman and Zhou (2021) document 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.[3] The coordinator

[2]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*.

[3]Townships are geographic partitions of the entire county. On average, each home visitor is in charge of 8 households' home visits.

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 work of the home visitors.

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 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 once a month and record monthly observations on the caregiver, the child, and home visitor and their interactions.

The visitors engage with households weekly and provide one hour parenting or caregiving guidance and support based on the Jamaica program protocols.[4] In 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, who played with the child after the visit and with what frequency), and child performance (e.g., the tasks taught in the last week, and new tasks in the current week). Heckman and Zhou (2021) document 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

The randomized control trial we study is based on a village (cluster) level matched-pair design. Bai (2019) shows that this design is optimal for minimizing the mean-squared

[4]The protocols are based on those used by the Jamaica program, adapted to Chinese culture (e.g., changing songs into popular Chinese songs, adding the background of pictures, which are familiar to Chinese people). The protocol for children younger than 18 months old focuses on motor and language skill training. After 18 months old, the protocol adds more cognitive skill content (e.g., classification, pairing, and picture puzzles).

error of estimates of average treatment effects. Implementation is in three steps. First, we examine the entire universe of eligible villages in Huachi county.[5] Second, based on both household surveys and village-level administrative data, the similarities of villages are assessed using a Mahalanobis metric of resident and village characteristics. To minimize the Mahalanobis metric in each pair, we sort the villages by the metric and pair the closest using the nonparametric belief propagation (nbp) matching method.[6]

After matching village pairs, we randomly select one village within the pair into the treatment group and the other village into the control group.[7] Figure A2 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.[8]

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 the detailed child development assessment task measures.[9,10,11]

[5]The pre-treatment village-level covariates used for the matching village pairs include the: (1) "closeness with children" scores on the Home Observation for Measurement of the Environment Inventory (HOME IT) scale; (2) language skill scores on the HOME IT scale; (3) learning materials score on the HOME IT scale; (4) take-up rate of a nutrition supplement program in the village; (5) compliance rate for a county-wide nutrition program in the village; (6) percentage of left-behind children in the children sample; (7) per capita net income in the village; (8) 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.

[6]Lu et al. (2011)

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

[8]Appendix B documents baseline comparisons.

[9]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 observation rather than a parental report. It is an inventory of 125 tasks including four aspects of skill measures: personal-social (getting along with people and caring for personal needs), fine motor-adaptive (eye hand 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). See Appendix B for more details on the test.

[10]Appendix C gives both the English and Chinese versions of the Denver II Test measure tables.

[11]The Bayley III test converts composite scores into scaled scores based on age, which are more useful in clinical practice. However, using itemized Denver II test measures, it is also possible to achieve the same

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 average treatment effects on unweighted sums of item scores within each category. Item scores are binary indicators of performance on a task. We use robust statistical methods to adjust for missing data and allow disturbances within villages to be correlated, analyzing treatment effects on the proportion of items passed in the Denver test by each skill category at both the county and village level (Cameron et al., 2008).

A major drawback to evaluating average treatment effects in the standard fashion, by using the proportion of items correctly answered, is that it 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 (van der Linden, 2016) and recover *individual* latent skills that generate item responses. We identify experimentally-induced improvements in latent skills and also improvements in utilization of skills to perform item-specific tasks.

goal. The Bayley III test targets infants and children between 1-42 months old and includes the examiner's observation (cognitive, motor, and language skills) and parent 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.

3.1 County level Average Treatment Effects

It is helpful for our exposition to define some notation. The universe of villages is $\{1, \ldots, Y\}$. 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 $\overline{\boldsymbol{Z}}(v)$. Closeness is calibrated by a Mahalanobis metric:

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

where $\sum$ is the covariance matrix of Z computed over all villages.

A coin is tossed to determine which village of (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 simple aggregates of standardized scores estimated from the following specification:

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

where Y_{iv}^j is the standardized scores for outcome j 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. $Y_{iv}^j = D_{v(i)}Y_{iv}^j(1) + (1 - D_{v(i)})Y_{iv}^j(0)$ where $Y_{iv}^j(d)$ denotes the vector of outcomes fixing treatment status d. The treatment assignment design implies that

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

Define the full array of right hand side variables in (1) as $\boldsymbol{X_{iv}}$.

Treatment is at the village level. We allow the idiosyncratic shock term ε_{iv} for child i to be arbitrarily correlated with $\varepsilon_{i'v}$ for any other child $i' \neq i$ in the same village v, but the idiosyncratic shocks are assumed to be independent across villages, i.e. $\varepsilon^j_{iv} \perp\!\!\!\perp \varepsilon^j_{kv'}$ for $\forall i \in v$ and $\forall k \in v', v \neq v'$. Residual plots 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 $v \neq v'$.[12]

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 the ratio of clusters to observations in the cluster converges to a constant. This is true if β^j_1 is constant across people or varies across people, although different parameters are identified if β^j_1 depends on the treatment effect.

However, the standard cluster-robust variance estimator (CRVE), $(\boldsymbol{X'X})^{-1}(\sum_{v=1}^{V} \boldsymbol{X'_v}\boldsymbol{\hat{\Omega}_v}\boldsymbol{X_v})(\boldsymbol{X'X})^{-1}$, is biased when $\hat{\Omega}_v$ is estimated by using the OLS residuals $\hat{\epsilon}_v$: $E(\boldsymbol{\hat{\epsilon}_v}\boldsymbol{\hat{\epsilon}'_v})$.[13] The bias depends on the form of $\boldsymbol{\Omega_v}$. Cameron et al. (2008) discuss this problem and show that the wild cluster bootstrap has good performance for making cluster-robust inferences. Details of the wild bootstrap procedures we use are presented in Appendix F.[14]

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).[15,16]

[12] $\boldsymbol{X_v}$ indicates $\boldsymbol{X}$ in the v^{th} 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 the indicators of the matched pair.

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

[14] Since we have 55 clusters, recent concerns raised about the wild bootstrap do not apply. See Canay et al. (2019).

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

[16] Appendix D documents the details of 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.

Table 2 presents the treatment effects for each skill category using standardized outcome measures.[17,18] Columns (1), (2), and (4) use all available data samples, and columns (3) and (5) only use samples of children who are younger than 2 years old at the time in September of 2015 when the program started. The treated younger children have at least one year of exposure to the intervention.[19]

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.[20] On average, treated children's scores are 0.7 standard deviation higher than those in the control group. In the first row, we see that at midline (about 9 months after 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 control group. At the end of the intervention, treatment effects on language and cognitive skills have effect sizes greater than 1.1. The intervention significantly improves the treated children's language and cognitive skills. The magnitude of the age-adjusted treatment effects increases when the children in the treatment group have longer exposures to home visitors (see columns (3) and (5)).

The intervention significantly improves social-emotional skills at midline, fine motor skills at the end of the intervention, and produces no significant improvement in gross motor skills. This finding is consistent with the design of the curriculum which focuses more on language and cognitive skill development.[21,22]

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

[18] 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.

[19] 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; and (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.

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

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

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

Tables 3-4 display the county level treatment effects by gender. An interesting finding, consistent with recurrent findings in literature (Elango et al., 2016), is that the intervention improves boys' language and cognitive skills much more than those of girls. At midline, the treatment effect size for girls is 0.4, and 0.9 for boys, respectively. At the end of the intervention, the effect size is about 0.9 for the girls and 1.1 for the boys. One reason for this is a threshold effect: on average girls are relatively more developed than boys at the same age in early childhood. The girls in the treatment group also have better performance in terms of social-emotional skills.[23]

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

Table 2: Treatment Effects on Standardized Scores

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

Notes: 1. 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.

Table 3: Treatment Effects on Standardized Scores

(Female)

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

Notes: 1. 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.

Table 4: Treatment Effects on Standardized Scores

(Male)

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

Notes: 1. 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.

Appendix G adds variables measuring interactions between home visitor and caregiver, and home visitor and child, and a variable capturing home visitor teaching ability.[24] The only strong pattern that emerges is that good caregiver–home visitor interactions promote language and cognitive skills. This may be a consequence of the uniform quality of the home visitors, but the evidence speaks otherwise.[25]

Appendix I reports village-level treatment effects. Many are not statistically different from zero. At the extremes of the distribution and accounting for the ordering, some estimated treatment effects for cognition are strongly statistically significant at endline. In those villages, interaction quality assessed by supervisors is higher.

3.2 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 of performing it 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.[26] 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 mapping between skills and item scores (i.e., whether treated children better utilize their skills). Shifts in these mappings can be due to improvements in children's ability to utilize skills.

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

[25] Table G2 in Appendix G shows the considerable dispersion in these measures.

[26] In the data, for each individual, we have more than 70 items per skill to measure task performances on the Denver test.

3.2.1 Model Specification

The outcomes we study are children's performances on individual tasks measured by performance on items on a test. There are N_J 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, $\boldsymbol{\theta}$.

Let $\tilde{Y}_i^{j,k}(d)$ be the binary-valued outcome variable indicating mastery of task j in 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 to the determinants on outcome on task j as

$$\tilde{Y}_i^{j,k}(d) = \boldsymbol{X_i'\beta^d} + \delta^j + \left(\boldsymbol{\theta_i^d}\right)' \boldsymbol{\alpha^{j,d}} + \varepsilon_i^j, \quad j = 1, N_J. \tag{3}$$

$$\tilde{Y}_i^{j,k} = \begin{cases} 1 & \tilde{Y}_i^{j,k} > 0 \\ 0 & \tilde{Y}_i^{j,k} \leq 0 \end{cases}$$

where $\boldsymbol{\alpha^{j,d}}$ is an array of factor loadings, δ^j is a task difficulty parameter and the coefficients $\boldsymbol{\beta^d}$, $\boldsymbol{\alpha^{j,d}}$ may depend on treatment as well as the latent skills.

This model interprets the intervention as shaping skills that are mapped into performances on tasks. An alternative interpretation is that the $\boldsymbol{\alpha^{j,d}}$ parameters are enhancements of skill. The intervention shifts $\boldsymbol{\alpha^{j,d}}$. Thus, $(\boldsymbol{\theta_i^d})'\boldsymbol{\alpha^{j,d}}$ is a bundle of effective skills from intervention $D = d$.

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 ε_i^j is unit normal, independent of the other right hand-side variables. This data has the structure of a panel, except 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.

Fernández-Val and Weidner (2016) show that estimators of the model are 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. These conditions apply in our sample with large numbers of tasks per person and observations. Factor models require normalizations if we seek to isolate θ^d from $\alpha^{j,d}$. Since $\boldsymbol{\theta_i^{d\prime}\alpha^{j,d}} = \left(\boldsymbol{\theta_i^d}\right)' \boldsymbol{AA^{-1}\alpha^{j,d}}$, the factors and factor loadings are intrinsically arbitrary unless some scale is set. We can avoid such normalizations if we are content to measure the shifts in effective skills, $\boldsymbol{\theta_i^{d\prime}\alpha^{j,d}}$. 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 $\boldsymbol{\alpha^{j,d}}$. We report estimates for $\boldsymbol{\theta_i^d}$ and $\boldsymbol{\alpha^{j,d}}$ separately and then as a bundle of effective skills $\left(\boldsymbol{\theta_i^d}\right)' \boldsymbol{\alpha^{j,d}}$.

Following traditions in the Rasch model literature (van der Linden, 2016), we assume that $\delta^{j,k}$ is an invariant task difficulty parameter intrinsic to the measurement system and independent of treatment status. This assumes 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. To interpret the factors, we assume that performance on K of N_J tasks $(K \leq N_J)$ depends only on one factor, what Cunha et al. (2010) call the "dedicated factor case." We generalize their analysis by requiring only that a subset of tasks are dedicated for any measurement of skills. We normalize the factor loading matrix so the first K rows form an $\boldsymbol{I_{K,K}}$ identity matrix. For $K = 4$ items, we assume that they load on one skill.[27] The remaining factor loading matrix for the vector of N_J outcome is unrestricted:

[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.

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

We report sensitivity analyses using a variety of plausible normalizations in Appendix L. We find that the estimates of $\boldsymbol{\alpha}^{j,d}$ 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 et al. (2021) to estimate panel probit models with multiple latent skill factors.[29]

3.2.2 Estimates

Table 5 presents estimates of $\boldsymbol{\beta}^d$. 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 language and cognitive task items between our model estimates and the data. We also fit the data well with the other types of tasks.[30]

[28] 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.
[29] Details regarding the method are presented in Appendix J.
[30] See Appendix K.

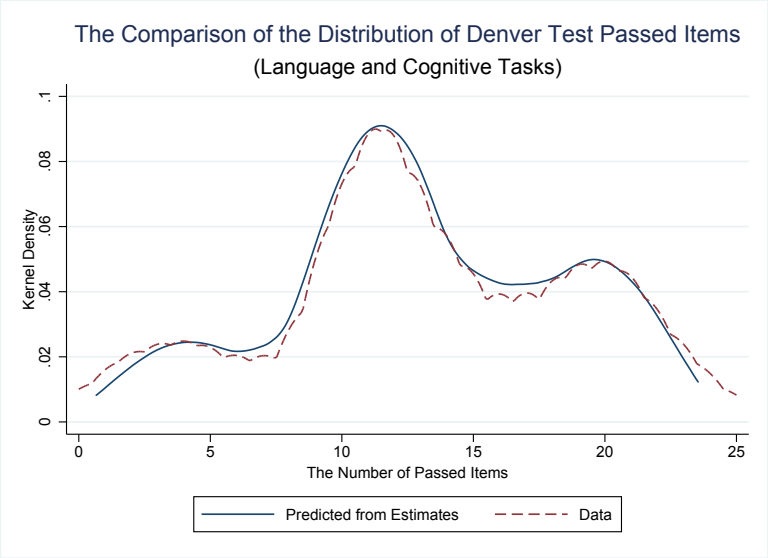


Figure 2: The Distribution of Denver Test Passed Items

Figure 3 shows the array of estimated difficulty level parameters δ^j 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 level of difficulty with later items. The difficulty level parameters δ^j 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 5: 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.

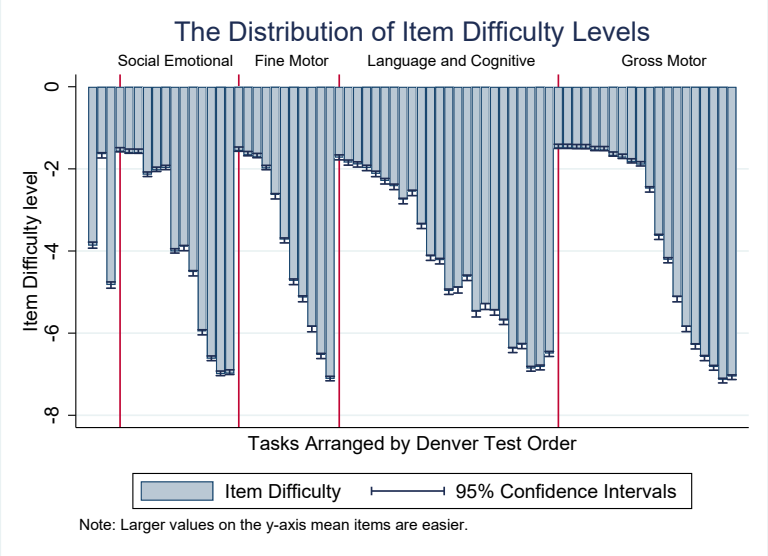
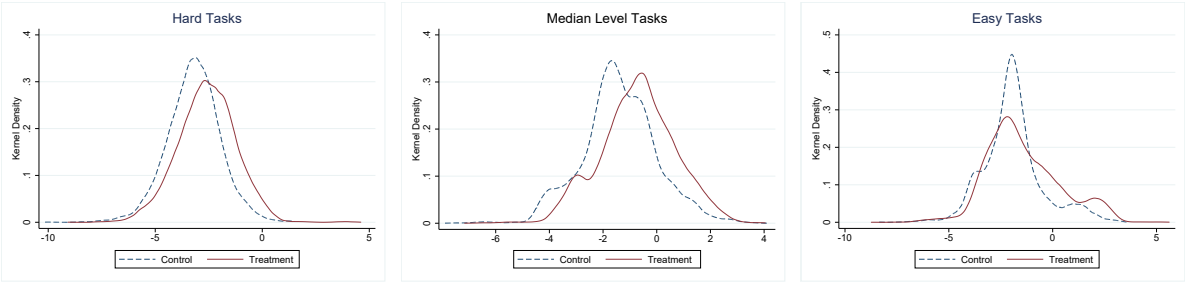


Figure 3: The Distribution of Denver Task Item Difficulty Levels

One advantage of our approach is that we can estimate individual level latent skill factors. First, Table 6 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 7 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 Distribution of $(\boldsymbol{\alpha}^{j'}\boldsymbol{\theta}_i)^{\dagger}$

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

Figure 4 plots the products of estimated skill factor loadings $\boldsymbol{\alpha}^j$ and the latent skill factors based on the Denver task difficulty levels.[31] The loadings for the treatment group are larger for the harder tasks and medium tasks, but smaller for easier tasks, which indicates that the easier tasks are not helpful for detecting treatment effects on child skill development. The loadings have similar patterns across treatment and the control groups for other skills. Estimates of aggregates of loadings are precisely estimated and for most tasks, we reject the hypothesis that $\alpha^{j,1} = \alpha^{j,0}$.[32] The only strong correlations are those between socioemotional skills and fine motor skills.

[31] 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 median level tasks. All the ranks are based on the estimates of the task difficulty level parameters.

[32] In Appendix L, Tables L3-L4 provide the tables for item-by-item tests. Social-emotional item loadings are not precisely estimated.

Table 6: 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. 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 7: 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

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

Table 8: Skill Loadings on Denver Test Tasks (α^j) Latent Skills

Control			Treatment			p-value
Skill Loadings	Mean	S. D.	Skill Loadings	Mean	S.D.	
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 $\alpha^{j,0}$ and $\alpha^{j,1}$, 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. Table 8 gives the summary statistics 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 importance of considering the task difficulty parameters in the model, in this section, 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 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 9. Notice that the estimates of a model without task difficulty parameters are very different from the estimates with the difficulty parameters. Such a model 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.

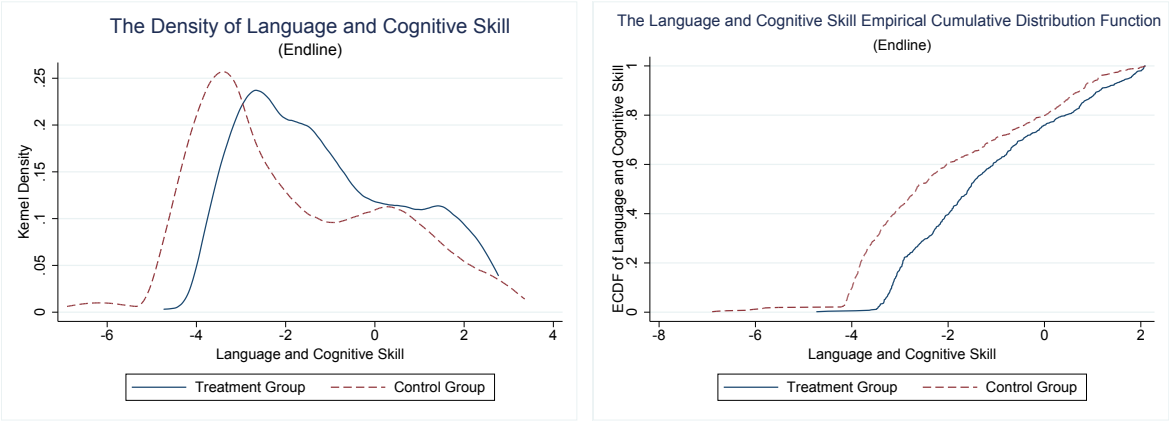
Table 9: Comparing Treatment Effects Based on Two Models With and Without Difficulty Parameters

	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. 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

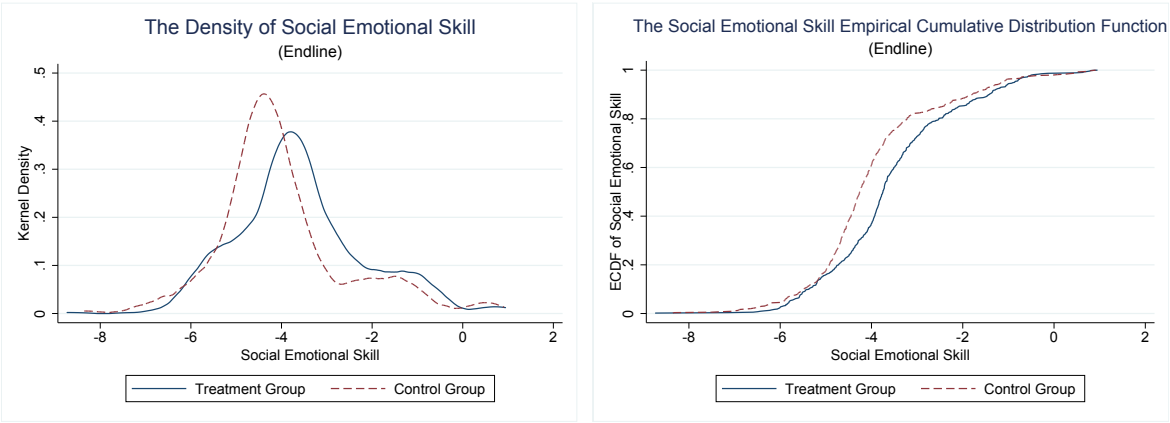
We compare the language skill distributions of the control and treatment groups. Figure 5 (a) shows that the density of language and cognitive skills for the treatment group shifts right and also has a fatter upper tail than the one in the control group. Figure 5 (b) shows that at almost every point of the cumulative distribution, language and cognitive skills are larger in the treated group than in the control group.



(a) (b)

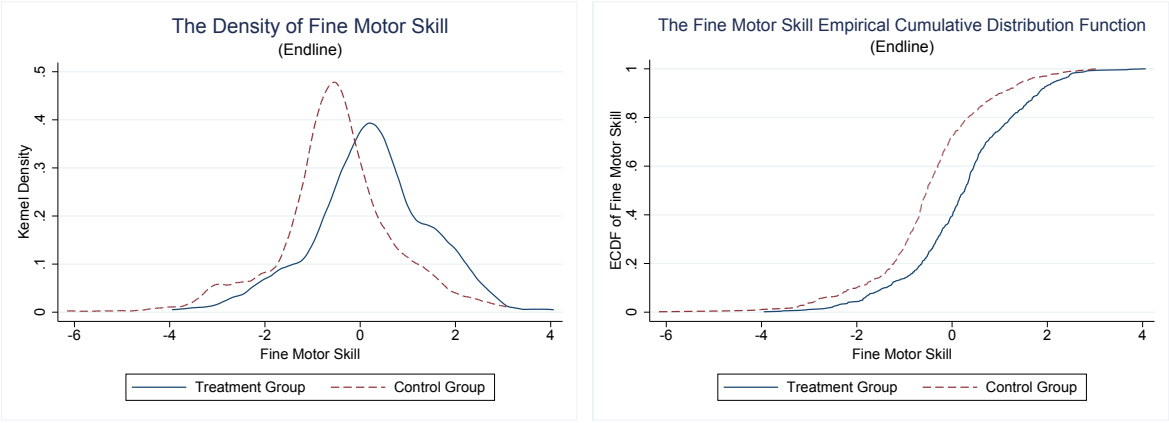
Figure 5: Language and Cognitive Skills Distribution

Switching focus to social-emotional and fine motor skills, Figures 6 (a) and 7 (a) present the densities of social-emotional and fine motor skills. Children in the treatment group are more concentrated at the upper ends of the distributions.



(a) (b)

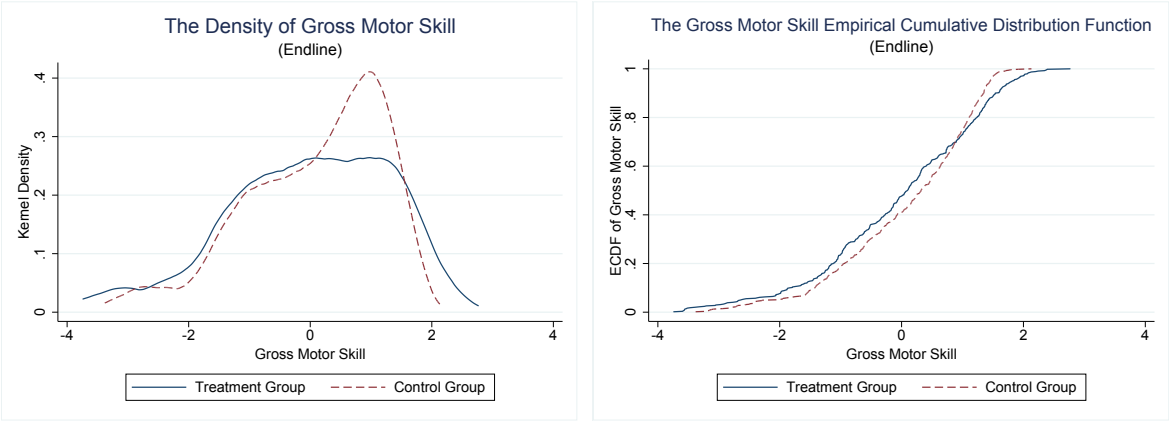
Figure 6: Social-emotional Skills Distribution



(a) (b)

Figure 7: Fine Motor Skills Distribution

For gross motor skills, we find that the factor distributions are similar between the control and the treatment groups. Figures 8 (a) and (b) show that both the densities and CDFs of the two gross motor skills distributions are close. In summary, language and cognitive, social-emotional, and fine motor skills were substantially improved by the program. Notice that looking solely at mean treatment effects, we only find significant improvement in language and cognitive skills and not strong effects on fine motor and social-emotional skills by the end of the intervention. Mean treatment effects are the combination of mean latent skills multiplied by skill loadings, which play a minor role as we show in the next section.



(a) (b)

Figure 8: Gross Motor Skills Distribution

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 Source of Treatment Effects

Average treatment effects produced from the experiment can arise 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 Denver test item j, the outcome measured is as follows:

$$y_i^{j*} = \boldsymbol{X_i'}(D_i\boldsymbol{\beta^{j,1}} + (1-D_i)\boldsymbol{\beta^{j,0}}) + D_i((\boldsymbol{\theta_i^1})'\boldsymbol{\gamma^{j,1}}) + (1-D_i)((\boldsymbol{\theta_i^0})'\boldsymbol{\gamma^{j,0}}) + \varepsilon_i^j \tag{5}$$

where, $\tilde{Y}_i^j$ is latent task measure j in the Denver test, $\boldsymbol{\theta_i^d}$ is child i's latent skill vector, and $\boldsymbol{\gamma^{j,d}}$ is the latent skill loading vector. D_i is the indicator of treatment status. We assume

that $\varepsilon_i^j \perp\!\!\!\perp \boldsymbol{X}$ and $\boldsymbol{\theta_i^d}, \forall\ j \in \{1, \ldots, N_J\}$ and $\varepsilon_i^j \perp\!\!\!\perp \varepsilon_i^m, \forall\ m, j \in \{1, \ldots, N_J\}, m \neq j$

$$\underbrace{\sum_{j \in \{1,\ldots,N_J\}} \tilde{Y}_i^j}_{\text{Denver Test Score } Y_i} = \sum_{j \in \{1,\ldots,N_J\}} \boldsymbol{X_i'}(D_i \boldsymbol{\beta^{j,1}} + (1 - D_i)\boldsymbol{\beta^{j,0}}) + D_i(\sum_{j \in \{1,\ldots,N_J\}} (\boldsymbol{\theta_i^1})' \boldsymbol{\alpha^{j,1}}) \tag{6}$$

$$+(1 - D_i)(\sum_{j \in \{1,\ldots,N_J\}} (\boldsymbol{\theta_i^0})' \boldsymbol{\alpha^{j,0}}) + \sum_{j \in \{1,\ldots,N_J\}} \varepsilon_i^j .$$

Define $\tilde{\lambda}$ as the mean difference in the latent skills produced by the intervention:

$$\tilde{\lambda} := E(\sum_{j \in \{1,\ldots,N_J\}} (\boldsymbol{\theta_i^1})' \boldsymbol{\gamma}^{j,1} | \boldsymbol{x_i}, D_i = 1) - E(\sum_{j \in \{1,\ldots,N_J\}} (\boldsymbol{\theta_i^0})' \boldsymbol{\gamma}^{j,0} | \boldsymbol{x_i}, D_i = 0). \tag{7}$$

Since we recover the individual latent skills $\boldsymbol{\theta_i^d}$, equation (7) provides another way to evaluate the average treatment effects on Denver test scores. We compare the treatment effects obtained from the experiment with the estimates based on our model of latent skills in Table 10.

The point estimates of the average treatment effects are almost identical using these two methods. We cannot reject the hypothesis that the two estimates are the same.

4.2 Decomposing Treatment Effects

Experimental treatment effects may not only arise from enhancements of latent skills but also from changes in the mapping from skills to tasks. In order to understand the source home visiting intervention treatment effects, in this section, 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 skill factors.

For each item j, the outcome Y_i^j is:

$$Y_i^{j,d} = 1(\boldsymbol{X_i'} \boldsymbol{\beta^{j,d}} + \delta^j + (\boldsymbol{\theta_i^d})' \boldsymbol{\alpha^{j,d}} + \varepsilon_i^j > 0) \tag{8}$$

where we assume $\varepsilon_i^j \sim N(0, 1)$. Home visiting treatment effects come from three chan-

Table 10
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. 95% confidence intervals in brackets are constructed by wild bootstrap clustered at the village level.
2. The ATE effect from the OLS model is based on equation (1) and the ATE effect from the factor model is based on equation (7).
3. The ATE estimates reported in this table are conditional on the pre-treatment covariates, which are consistent with Table 2 column (5).
4. 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.

nels: changes in the observable coefficient, $\boldsymbol{\beta}^{j,d}$, changes in 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 distribution of $(\boldsymbol{\theta^1}, \boldsymbol{X})$ and $(\boldsymbol{\theta^0}, \boldsymbol{X})$ in the treatment and control populations, respectively. The population treatment effect for item j can be decomposed as follows:

$$
\begin{aligned}
\Pr(Y^{j,1}=1) - \Pr(Y^{j,0}=1) \quad = \quad & \underbrace{\int \{\Phi([\boldsymbol{x'\beta}^{j,\mathbf{1}} + \delta^j + (\boldsymbol{\theta^1})'\boldsymbol{\alpha}^{j,\mathbf{1}}]) - \Phi([\boldsymbol{x'\beta}^{j,\mathbf{0}} + \delta^j + (\boldsymbol{\theta^1})'\boldsymbol{\alpha}^{j,\mathbf{1}}])\} dF^1(\boldsymbol{\theta^1}, \boldsymbol{X})}_{\text{From Estimated Coefficients of X}} \\
+ \quad & \underbrace{\int \{\Phi([\boldsymbol{x'\beta}^{j,\mathbf{0}} + \delta^j + (\boldsymbol{\theta^1})'\boldsymbol{\alpha}^{j,\mathbf{1}}]) - \Phi([\boldsymbol{x'\beta}^{j,\mathbf{0}} + \delta^j + (\boldsymbol{\theta^1})'\boldsymbol{\alpha}^{j,\mathbf{0}}])\} dF^1(\boldsymbol{\theta^1}, \boldsymbol{X})}_{\text{From Latent Skill Loadings}} \\
+ \quad & \underbrace{\int \Phi([\boldsymbol{x'\beta}^{j,\mathbf{0}} + \delta^j + (\boldsymbol{\theta^1})'\boldsymbol{\alpha}^{j,\mathbf{0}}]) dF^1(\boldsymbol{\theta^1}, \boldsymbol{X}) - \int \Phi([\boldsymbol{x'\beta}^{j,\mathbf{0}} + (\boldsymbol{\theta^0})'\boldsymbol{\alpha}^{j,\mathbf{0}}]) dF^0(\boldsymbol{\theta^0}, \boldsymbol{X})}_{\text{From Latent Skill Factors}}.
\end{aligned}
\tag{9}
$$

Notice that equation (9) holds over 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.[33] Table 11 reports the decomposition of treatment effects. The main drivers of treatment effects are increases in latent skills. The contributions from experimentally-induced changes in $\boldsymbol{\alpha}$ are not precisely estimated.

Table 11: Sources of the Treatment Effects

Tasks	Total Net Treatment Effects	From Observable Covariates	From Skill Loadings α	From Latent Skills θ
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. The total treatment effects are defined as $T_k = \sum_{j \in k}(\sum_{i \in D^1} \mathbf{1}^{j,\mathbf{1}} - \sum_{i \in D^0} \mathbf{1}^{j,\mathbf{0}})$

2. To make sure the observed covariates balance between treatment and control groups, we consider the sample which is younger than 46 months old and older than 12 months old.

3. Standard errors are reported in the parentheses.

5 Comparison of Growth Trajectories with the Source Program: Jamaica Reach Up and Learn

Table 12 shows that for comparable outcome measures at the early ages, and Figure 9 shows that the growth of cognitive skills in China REACH is on track with Jamaica Reach Up and Learn, which has been shown to generate substantial lifetime benefits (see Grantham-McGregor and Smith, 2016; Gertler et al., 2014). Treatment effects are comparable and we cannot reject the hypothesis that the treatment effects are the same across these two interventions.

[33] 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: Treatment Effects on China REACH and Jamaica Reach Up and Learn

	Panel A: China REACH Latent Skill Factors (after 21 Months' 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' Intervention)			
	Performance	Fine Motor	Hearing & 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

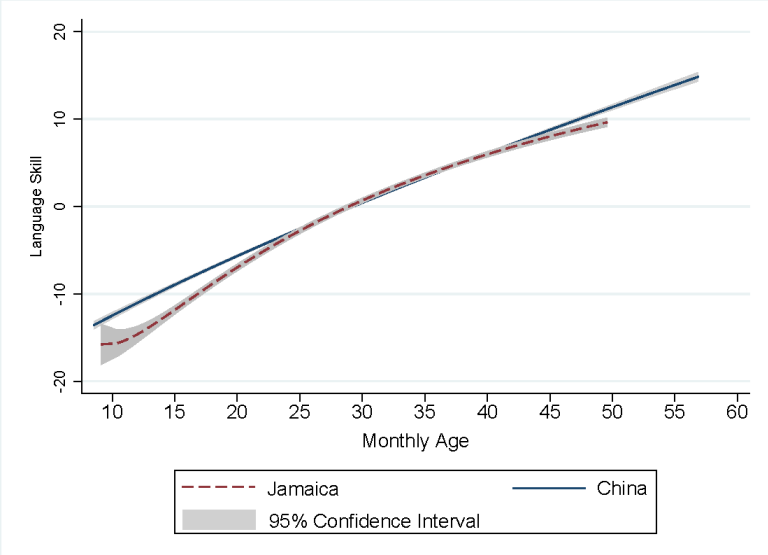
Notes: 1. For the China REACH program, 95% confidence intervals in brackets are constructed by wild bootstrap clustered at the village level.

2. For Jamaica Reach Up and Learn program, 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 are with respect to the null of equality of treatment effects across the programs.

Figure 9: Language Skill Growth Curve Comparison



6 Conclusion

This paper estimates the impacts from a large scale early childhood home visiting intervention program (China REACH) on child skill development that 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 provide a framework for understanding the mechanisms generating the standard treatment effects on child skill development that adjusts for difficulty of the various tasks used to assess the program. The program significantly improves child language, fine motor, and social-emotional skills. Impacts are largest in the most disadvantaged communities, as measured by home environments. Latent skill improvements explain about 90% of the treatment effects on language and cognitive skill development. The program also shifts the technology mapping latent skills into treatment effects, although this source explains less than 10% of the estimated treatment effects on average and is mostly concentrated on language skills. The latter source is quantitatively small and not precisely determined. Our analysis offers a prototype for measuring latent skills using diverse outcome measures and adjusting for the difficulty inherent in tasks.

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. 3.2.1

Bai, Y. (2019). Optimality of matched-pair designs in randomized controlled trials. Unpublished manuscript, University of Chicago. 2.1.1

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. 3, 3.1

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*, 1–45. 14

Chen, M., I. Fernández-Val, and M. Weidner (2021). Nonlinear factor models for network and panel data. *Journal of Econometrics 220*(2), 296–324. 3.2.1

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. 3.2.1

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. 3.1

Fernández-Val, I. and M. Weidner (2016). Individual and time effects in nonlinear panel models with large n, t. *Journal of Econometrics 192*(1), 291–312. 3.2.1

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. 23

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. 2, 5

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). 1, 2, 5

Heckman, J. J. and G. Karapakula (2019). Intergenerational and intragenerational externalities of the Perry Preschool project. NBER Working Paper 25889. 23

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. 1

Heckman, J. J. and J. Zhou (2021). Interactions as investments. Unpublished. 2.1, 21

HomVEE (2020). Early childhood home visiting models: Reviewing evidence of effectiveness, 2011-2020. OPRE Report 2020-126. 1

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. 1

Lu, B., R. Greevy, X. Xu, and C. Beck (2011). Optimal nonbipartite matching and its statistical applications. *American Statistics 65*(1), 21–30. 6

Maasoumi, E. and L. Wang (2019). The gender gap between earnings distributions. *Journal of Political Economy 127*(5), 2438–2504. 15

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. 11

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer. 3.1

van der Linden, W. J. (2016). *Handbook of Item Response Theory: Volume 1: Models*. CRC Press. 3, 3.2, 3.2.1