HCEO WORKING PAPER SERIES

Treatment Effects and the Measurement of Skills in a Prototypical Home Visiting Program

James Heckman
Bei Liu
Mai Lu
Jin Zhou

Working Paper

2020-047

06/2020

HUMAN CAPITAL AND
ECONOMIC OPPORTUNITY
GLOBAL WORKING GROUP

The University of Chicago
1126 E. 59th Street Box 107
Chicago IL 60637

www.hceconomics.org

Treatment Effects and the Measurement of Skills in a Prototypical Home Visiting Program[*]

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

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

[2]China Development Research Foundation

June 3, 2020

[*]CEHD acknowledges support from the Institute for New Economic Thinking, and the National Institutes of Health (R37-HD065072). 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 or the National Bureau of Economic Research. 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, Yike Wang, and Fuyao Wang for fundamental work on program design, implementation, and data cleaning support of the China REACH study. 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.

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 90% 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 10% of the average estimated treatment effects.

JEL Codes: J13, Z18

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

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

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

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., Grantham-McGregor and Smith, 2016) and are relatively low cost compared to other early childhood programs. They place minimal demands on the training required of the visitors and on the infrastructure required to support them. 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-plus 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 a strong impact on language and cognitive skills, fine motor skills, and social-emotional skills. Impacts are especially strong in the most disadvantaged communities.

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. We decompose conventional treatment effects into induced improvements in latent skills and improvements in the technology mapping skills into performance on tasks. Treatment effects mainly arise from boosts in skills. At least 90% of the estimated treatment effects are due to changes in latent inputs with the rest attributable to improvements in technologies.

This paper proceeds as follows. Section 2 describes the program and places it in context as a scaled version of an influential pilot 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 summarizes our findings. Supporting material is reported in a web appendix: `http://cehd.uchicago.edu/china-reach_home-visiting_appendix`.

2 China REACH

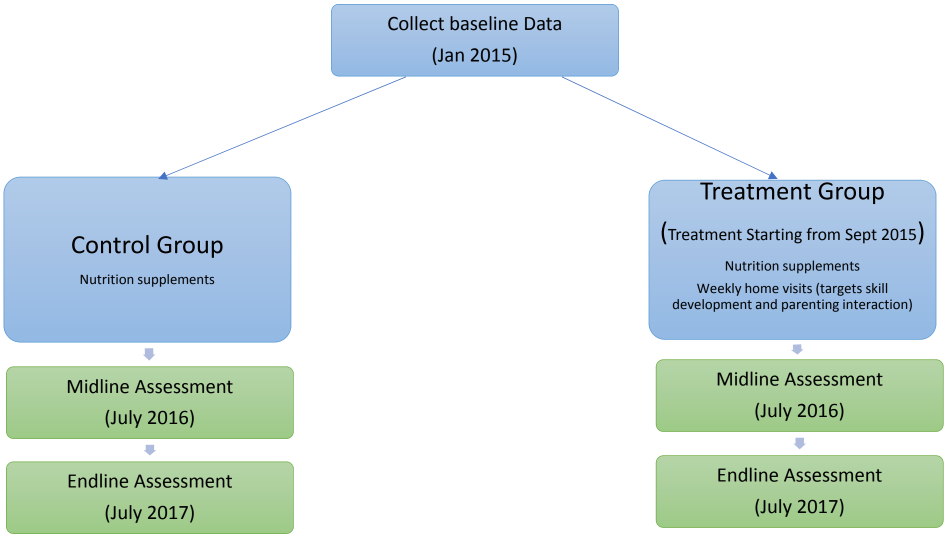
The ongoing Rural Education and Child Health (China REACH) project was initially launched in 2015 in response to a growing focus on, and call for, evidence-based pilot-to-policy analyses by China's State Council. China REACH is a large scale program evaluated by a randomized control trial (RCT) designed 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.[1] 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.

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

2.1 The Intervention Implemented

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



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 (2020) 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.[2] The coordinator prepares countywide training to oversee the township supervisors. The county project coordinator and township supervisors randomly attend home visits for spot checks to observe and review the 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 fu-

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

ture 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 of the caregiver, the child, and home visitor.

The visitors engage with households weekly and provide one hour parenting or caregiving guidance and support based on the Jamaica program protocols.[3] 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 (2020) document the content of China REACH curriculum. The curriculum includes more than 200 tasks related to language and cognitive skill development and has about 20 tasks targeting gross motor skills 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 error of estimates of treatment effects. Implementation is in three steps. First, we examine the entire universe of eligible villages in Huachi county.[4] Second, based on both house-

[3] 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).

[4] 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.

hold 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 with closest metric into pairs using the nonparametric belief propagation (nbp) matching method.[5]

After matching village pairs, we randomly select one within the pair into the treatment group and the other village into the control group.[6] 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.[7]

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.[8,9] 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 goal.[10]

[5] Lu et al. (2011)

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

[7] Appendix B documents baseline comparisons.

[8] The Denver II test is designed for clinicians, teachers, or early childhood professionals monitoring the development of infants and preschool age children. The test is primarily based on the examiner's actual 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.

[9] Appendix C gives both the English version and Chinese version Denver II Test measure tables.

[10] 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 better than the Bayley test in examining the delay of language development.

Table 1: China REACH Home Visiting Program Skill Content

Skill Category	Definition
Fine Motor	The skill of finger movements, such as grasping, releasing and stitching, drawing, and writing.
Gross Motor	This skill of 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 over 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).

There is, however, a major drawback to evaluating average treatment effects using the proportion of items correctly answered, although it is standard practice in the child development literature to do so. 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 use. We address this problem using a nonlinear measurement model 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.

3.1 County level Average Treatment Effects

We first report average treatment effects for simple aggregates of correct answers estimated from the following specification:

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

where Y_{iv}^{j} is outcome j for child i in village v, D_{iv} 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. Define the full array of right hand side variables in (1) as $\boldsymbol{X_{iv}}$. Let $Y_{iv}^{j}(d)$ denote the vector of outcomes fixing treatment status d. The treatment assignment design implies that

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

Bai et al. (2019) show that under model specification (1), idiosyncratic shocks ε_{iv} are independent at the individual level. Using their method, we can consistently estimate the average treatment effect. They also derive the asymptotic distribution of the estimated β_1^{j}.

As 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 $\varepsilon_{iv}^{j} \perp\!\!\!\perp \varepsilon_{kv'}^{j}$ 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 $E(\boldsymbol{\epsilon\epsilon'}) = \boldsymbol{\Omega}$ with V number of villages is block diagonal: $\boldsymbol{\Omega}_{vv'} = 0$; all $v \neq v'$.[11]

The standard cluster-robust variance estimator (CRVE), $(\boldsymbol{X'X})^{-1}(\sum_{v=1}^{V}\boldsymbol{X_v'}\boldsymbol{\hat{\Omega}_v}\boldsymbol{X_v})(\boldsymbol{X'X})^{-1}$,

[11] $\boldsymbol{X_v}$ 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.

is biased when $\hat{\Omega}_v$ is estimated by $E(\hat{\epsilon}_v\hat{\epsilon}_v')$.[12] The bias depends on the form of $\mathbf{\Omega}_v$. Cameron et al. (2008) discuss this problem and suggest that the wild cluster bootstrap has good performance for making cluster-robust inference. Details of the wild bootstrap procedures we use are presented in Appendix F.[13]

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

Table 2 presents the treatment effects for each skill category using standardized outcome measures.[16,17] 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 exposure to the intervention.[18]

The first row in Table 2 shows that the children in the treatment group are, on average, more likely to have higher language and cognitive skills.[19] On average, treated children's scores are 0.7 standard deviation higher than those in the control group. In the first row,

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

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

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

[15] 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.

[16] 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.

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

[18] There are two reasons for restricting the sample: (1) As claimed, we want the children in the treatment group to have enough exposure to the intervention; 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.

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

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 the control's. 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.[20],[21]

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.[22]

[20] Heckman and Zhou (2020) carefully document the intervention curriculum.

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

[22] This result is also found in the evaluation of the Perry Preschool Program (Heckman and Karapakula, 2019) and the Abecedarian preschool program (García 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 ≤ 2 Yrs at Enrollment	All	Children ≤ 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)

Denver Tasks	(1) All	(2) All	(3) Children $\leq$ 2 Yrs at Enrollment	(4) All	(5) Children $\leq$ 2 Yrs at Enrollment
			Midline		
Language and Cognitive	0.747***	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.

3.2 Village Level Treatment Effects

This section reports estimates of the village level treatment effects obtained from:

$$Y_i^j = \boldsymbol{Z}_i^{\boldsymbol{j}\prime}\boldsymbol{\beta} + \boldsymbol{Q}_v^{j\,\prime}\gamma_v^{j,p} + \varepsilon_i^j \tag{3}$$

where $\boldsymbol{Q}_v^j$ is a vector of dummy variables denoting village pairs in the county. Village level average treatment effects for the P villages in $Te_p^j, p \in \{1, \cdots, P\}$ are defined as follows:

$$Te_p^j = \gamma_{v'}^{j,p,D=1} - \gamma_v^{j,p,D=0}, \quad v', v \in p, \quad v' \neq v. \tag{4}$$

Figure 2 shows the distribution of village level treatments at midline ordered by magnitude.[23] For language and cognitive standardized scores, there are 40 villages with positive treatment effects, eight of which have significantly positive treatment effects. 17 villages' treatment effects have effect sizes greater than 1 standard deviation (relative to otherwise similar but untreated children). Six villages' treatment effects are greater than 2. In Appendix G, we adjust for cherry picking and show that the high-end outcomes are unlikely to arise by chance.

For fine motor scores, 31 villages have positive treatment effects; eight are statistically significant. In general, social-emotional treatment effects are not large: among 37 pairs with positive treatment effects, most of the treatment effects are statistically insignificant with only 2 pairs' treatment effects greater than 2. The distribution of gross motor skill village-level treatment effects are relatively symmetric: there were 25 pairs with positive effects and 30 with negative effects.

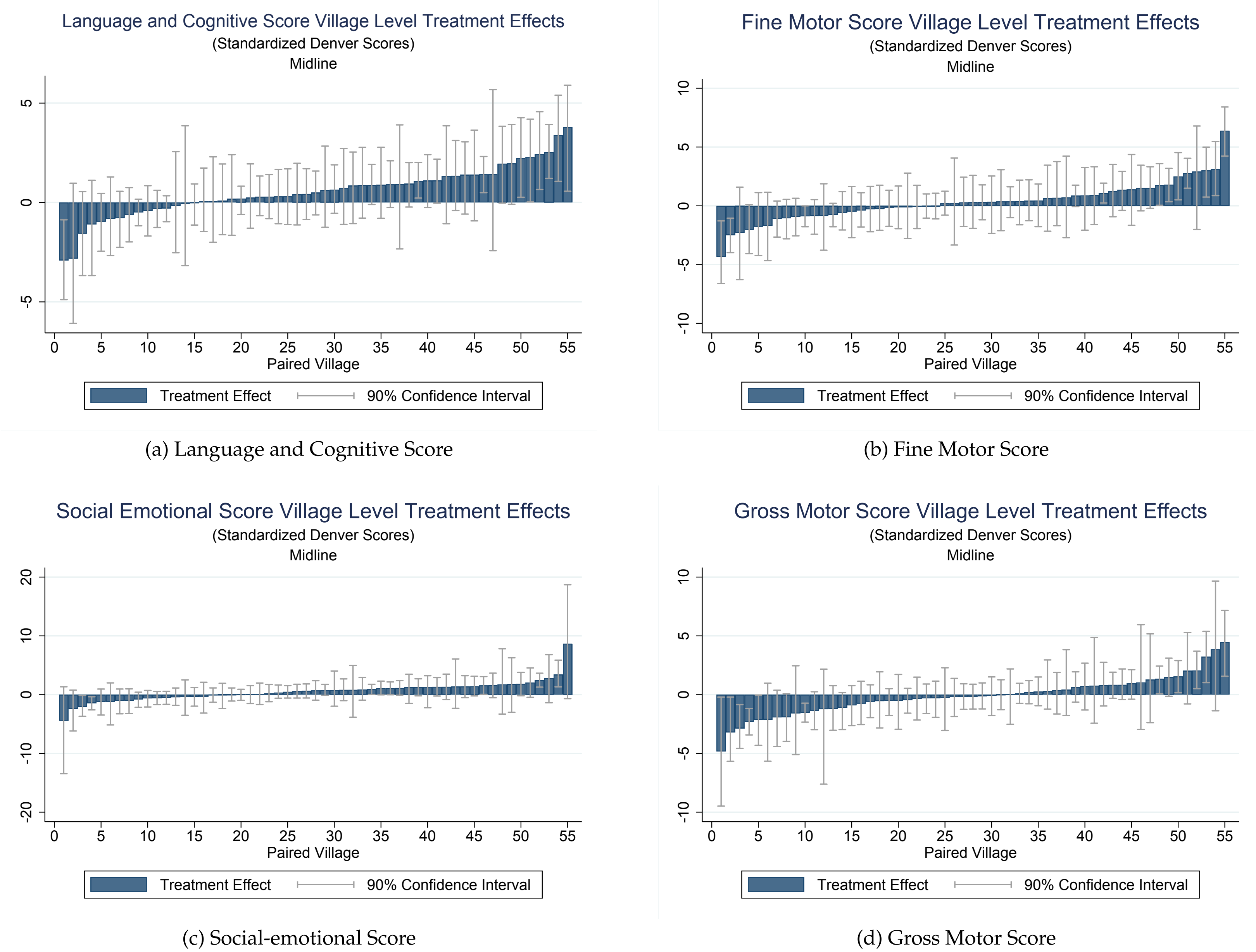
The village-level estimated treatment effects are consistent with the county-level estimates: the improvement in language and cognitive skills contributes most to the Denver test score treatment effects. For gross motor scores, the home visiting intervention has

[23] The confidence intervals are calculated using the wild bootstrap method.

little impact. Almost all estimated effects are statistically insignificant from zero.

Figure 3 gives the village level treatment effect distribution at the endline. (The identities of the village change across graphs which are ordered by magnitudes of treatment outcomes.) Language score treatment effects contribute the most to total score treatment effects: 41 pairs with positive treatment effects and seven of them are significant. The treatment effects at endline are greater than the ones at midline: 20 pairs of treatment effects are above 1, and seven pairs have effect sizes greater than 2. Treatment effects for social-emotional and gross motor skills are statistically significant. Tables 5 (midline) and 6 (endline) give raw correlations across village pairs of treatment effects across skills. At midline, the correlations are substantial: the language and gross motor skills correlate at about 0.38, language and social-emotional correlate at 0.42, and all other correlations are above 0.5. However, at the endline, only the two types of gross motor skills are highly correlated at the village level.

In Table 7, we compare the top ten treated villages with the highest language and cognitive treatment effects to those of the bottom ten treated villages. Before the home visiting intervention, except for the proportion of females children, there are no significant differences between the high-performing and low-performing villages. However, when we compare the measures evaluating home visit quality, we find that the qualities of interactions between the home visitor and caregiver/child are significantly better in the high-performing villages than those in the low-performing villages. Also, in the low-performing villages, the grandmother appears 1.6 times more often than in the high-performing villages. These effects of the quality of visitor interactions arise even though the home environments, on average, are worse in the high-performing villages than the low-performing villages before the intervention.



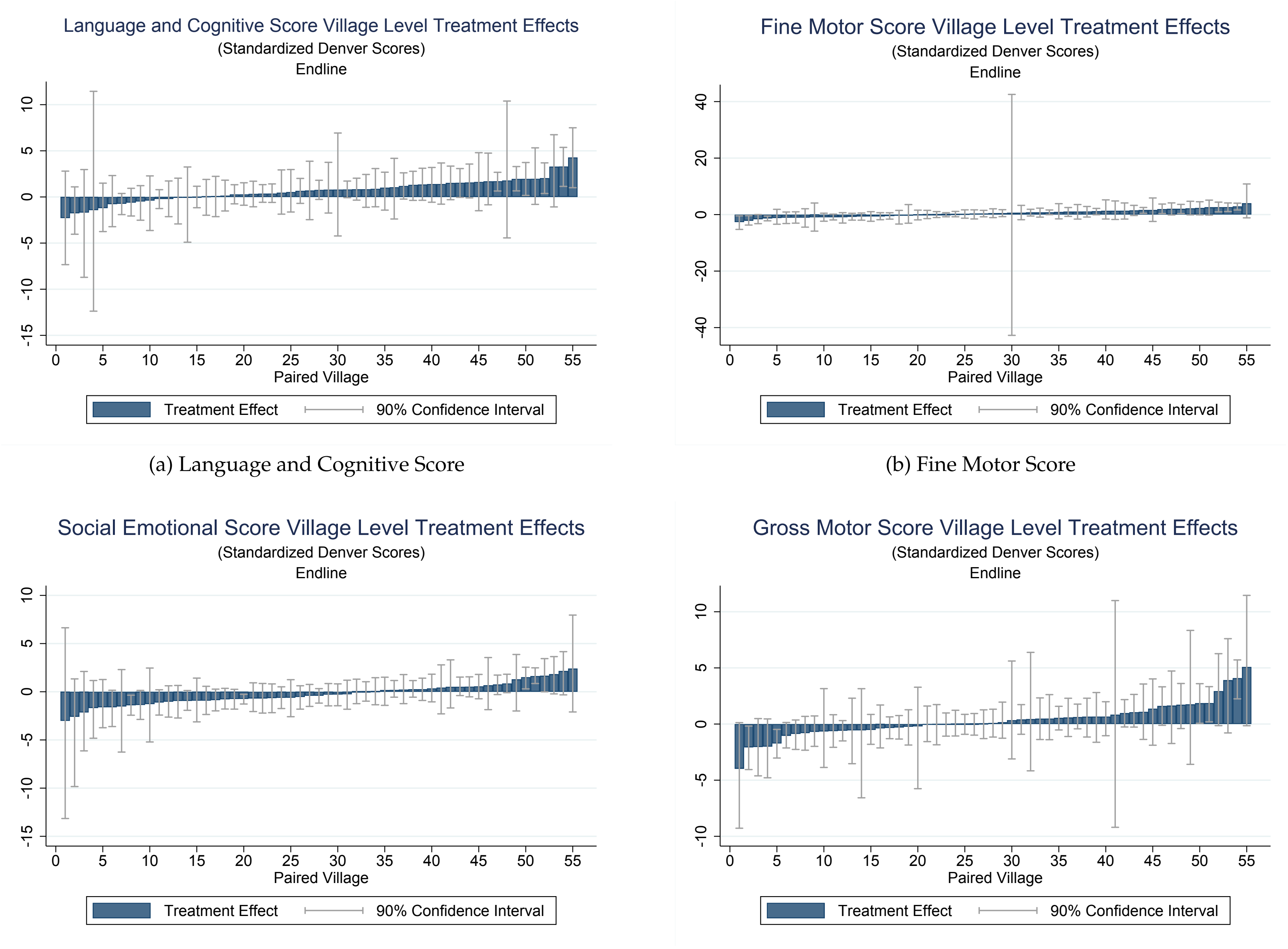
(a) Language and Cognitive Score

(b) Fine Motor Score

(c) Social-emotional Score

(d) Gross Motor Score

Figure 2: Village Level Treatment Effects at the Midline



(a) Language and Cognitive Score

(b) Fine Motor Score

(c) Social-emotional Score

(d) Gross Motor Score

Figure 3: Village Level Treatment Effects at the Endline

Table 5: Village Level Treatment Effect Rank Correlation (Midline)

	Social-emotional	Fine Motor	Language and Cognitive	Gross Motor
Social-emotional	1.00			
Fine Motor	0.59	1.00		
Language and Cognitive	0.42	0.50	1.00	
Gross Motor	0.55	0.55	0.38	1.00

Table 6: Village Level Treatment Effect Rank Correlation (Endline)

	Social-emotional	Fine Motor	Language and Cognitive	Gross Motor
Social-emotional	1.00			
Fine Motor	0.31	1.00		
Language and Cognitive	0.13	0.29	1.00	
Gross Motor	0.21	0.54	0.28	1.00

Table 7: The Comparison between High-performance Treated Villages and Low-performance Treated Villages

	High Performance Villages		Low Performance Villages		
	Mean	S.D.	Mean	S.D.	P-value
Monthly Age	37.55	5.43	37.41	5.45	0.95
Male Children	0.45	0.50	0.56	0.50	0.04
Years of Education					
Father	7.37	2.99	8.05	3.21	0.08
Mother	6.99	3.46	6.76	3.34	0.49
Grandmother	2.65	2.82	3.08	3.15	0.21
Baseline Measures					
Home Score					
Warmth	4.06	2.14	4.09	2.02	0.97
Verbal Skill	2.27	1.10	2.41	0.96	0.07
Hostility	0.16	0.46	0.11	0.37	0.11
Learning Materials	5.56	3.20	6.00	3.34	0.06
Outings	1.62	1.10	1.58	1.10	0.92
Willingness to Participate the Program	0.99	0.11	0.93	0.25	0.05
Household Income (Annual, RMB)	40984.2	5771.5	40118.9	8478.1	0.15
Measures During the Interventions					
Home Visitor Teaching Skill	-0.08	0.50	-0.11	0.57	0.04
The Quality of Interactions Between Home Visitor and Caregiver	-0.13	0.86	-0.25	1.01	0.00
The Quality of Interactions Between Home Visitor and Child	0.07	1.43	-0.14	1.36	0.00
Grandmother as the Caregiver During Home Visits	0.16	0.37	0.26	0.44	0.00

3.3 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 intrinsically uninterpretable, unless the difficulty is the same across tasks, which, is not true by design.

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.[24] 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.

3.3.1 Model Specification

The outcomes we study are children's performances on individual item tasks measured by performance above thresholds or correct answers on an item on a test. There are N_J tasks. We break down the tasks by skill categories (motor, cognitive, etc.) when we conduct our empirical analysis. Here we abstract from these categories to simplify notation. Performance on the item tasks is assumed to be generated, in part, by latent skills.

Let $Y^{j*}(d)$ denote a latent outcome for task 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. We write the mapping from latent skills to outcome j as

$$Y_i^{j*}(d) = \boldsymbol{X_i'}\boldsymbol{\beta^d} + \delta^j + (\boldsymbol{\theta_i^d})'\boldsymbol{\alpha^{j,d}} + \varepsilon_i^j \tag{5}$$

[24] In the data, for each individual, we have more than 70 items to measure his or her task performances in the Denver test.

$$Y_i^j = \begin{cases} 1 & Y_i^{j*} > 0 \\ 0 & Y_i^{j*} \leq 0 \end{cases}$$

where $\boldsymbol{\alpha}^{j,d}$ is a vector 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 conceptualizes the intervention as shaping a bundle of invariant skills that are mapped into performance on tasks. An alternative interpretation is that the $\boldsymbol{\alpha}^{j,d}$ parameters are enhancements of skill. The intervention $\boldsymbol{\alpha}^{j,d}$ so $(\boldsymbol{\theta}^d)'\boldsymbol{\alpha}^{j,d}$ is a bundle of effective skills from intervention $D = d$.

Using this model, under suitable normalizations we estimate the individual level latent skill factors $\boldsymbol{\theta}_i^d$, and not just the distribution of the latent skill factors, as in traditional models in psychology (see e.g., van der Linden, 2016). We assume that ε_i^j is unit normal, independent of the other right hand-side variables. For this panel probit model with latent skills, we estimate both the parameters of observed covariates, the latent factors, and the effects of latent skill factors on outcomes. Fernández-Val and Weidner (2016) show that the estimators are asymptotically unbiased when the number of observations $N_I \to \infty$ 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 and observations.

Factor models require normalizations. Since $\boldsymbol{\theta}_i^{d\prime}\boldsymbol{\alpha}^{j,d} = (\boldsymbol{\theta}_i^d)'\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{\alpha}^{j,d}$, the factors and factor loadings are intrinsically arbitrary unless a scale is set. Using a normalization applied by Anderson and Rubin (1956), we identify both the vector $\boldsymbol{\theta}_i^d$ and $\boldsymbol{\alpha}^{j,d}$. This issue is moot if we only seek to estimate effective skills, $(\boldsymbol{\theta}_i^d)'\boldsymbol{\alpha}^{j,d}$. We report estimates for $\boldsymbol{\theta}_i^d$ and $\boldsymbol{\alpha}^{j,d}$ separately and then as a bundle of effective skills.

Following traditions in the Rasch model literature, we assume that δ_j is an invariant task difficulty parameter intrinsic to the measurement system and independent of treatment status. This is one way to assure 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", except we only require 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 the $K = 4$ normalized items, we assume that they load on one skill.[25] After normalization, the factor loading matrix for the vector of N_J outcome is:

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

We report sensitivity analyses with a variety of plausible normalizations in Appendix J and find that the estimates of $\boldsymbol{\alpha}$ reported in the text are stable under different normalizations.[26] Our results are quantitatively robust. We use the estimation procedure proposed by Chen et al. (2019) to estimate panel probit models with multiple latent skill factors.[27]

[25] 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.

[26] In Appendix J (page 44), we compare the distribution of the skill loadings under different normalizations. We find that the results are robust when we choose the items within the median range difficulty level.

[27] Details regarding the method are presented in Appendix H.

3.3.2 Estimates

Table 8 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. In Figure 4, we compare the distribution of language and cognitive task items between our model estimates and the data. We fit the data well and as we do the other types of tasks.[28]

Figure 5 shows the array of difficulty level parameters δ^j for each task item. When the item difficulty level increases, the estimates become smaller. 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 very 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 failed to answer all hard 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 design of social-emotional task items could also be improved.

Table 8: Estimates 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]

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.

One advantage of our model is that we can examine individual level latent skill fac-

[28] See Appendix I.

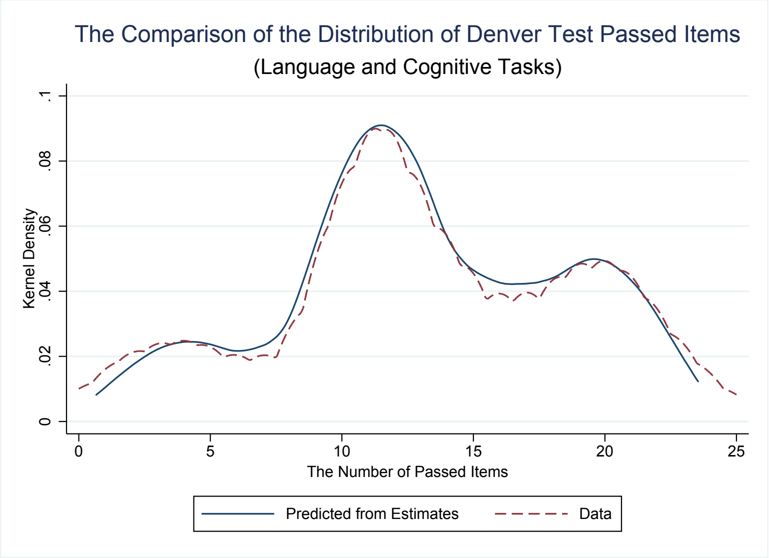


Figure 4: The Distribution of Denver Test Passed Items

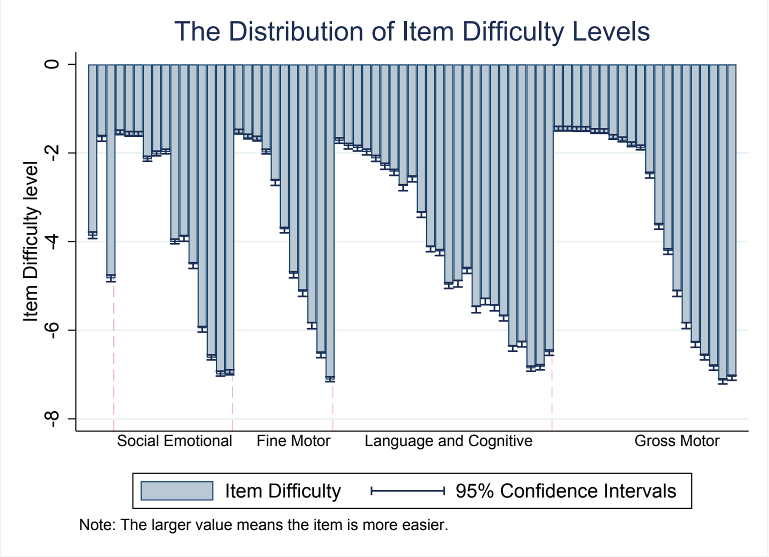
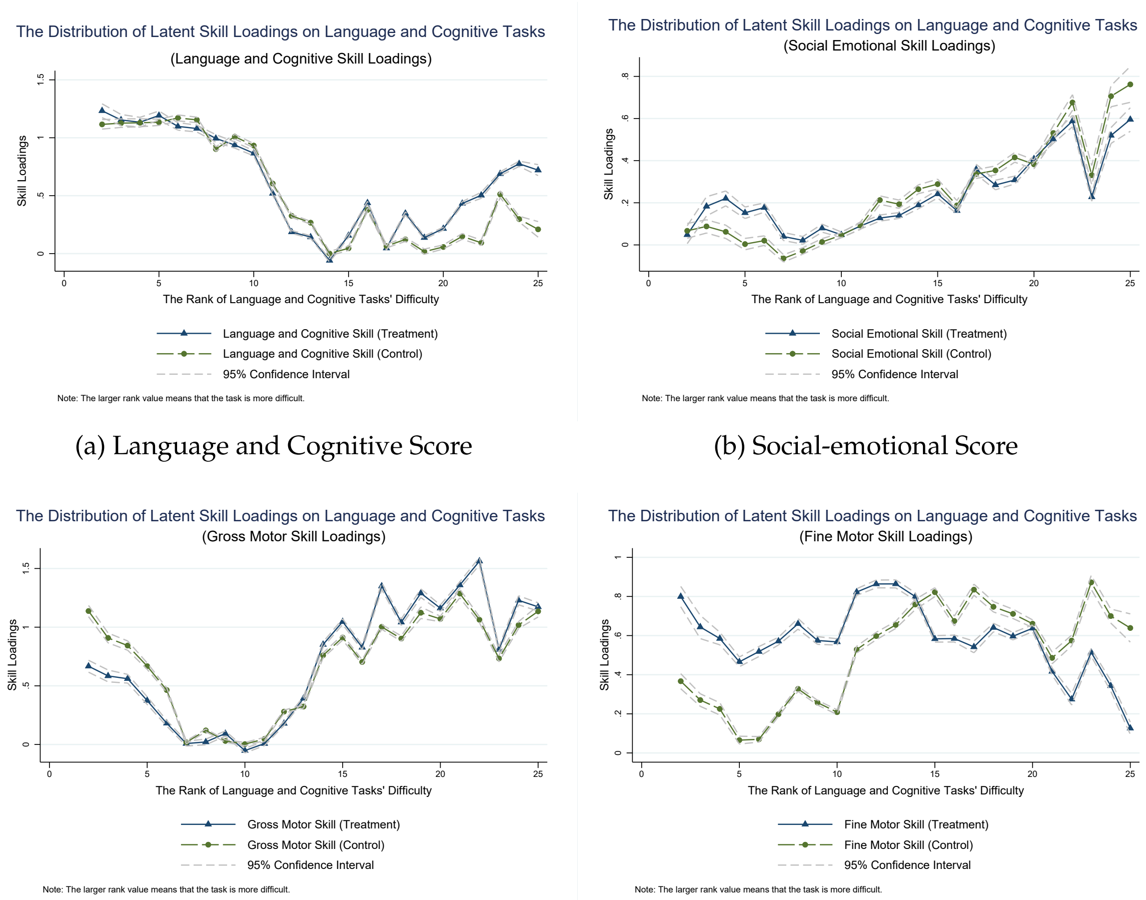


Figure 5: The Distribution of Denver Task Item Difficulty Levels

tors. Firstly, Table 9 presents the treatment effects for four latent skill factors. For these factors, these are the means of the estimated skill distributions. We find that except for gross motor skills, all other latent skill factors in the treatment group are significantly higher than those in the control group. Also, language and cognitive skills are negatively correlated with gross motor skills and positively correlated with social-emotional and fine motor skills. 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.



(a) Language and Cognitive Score

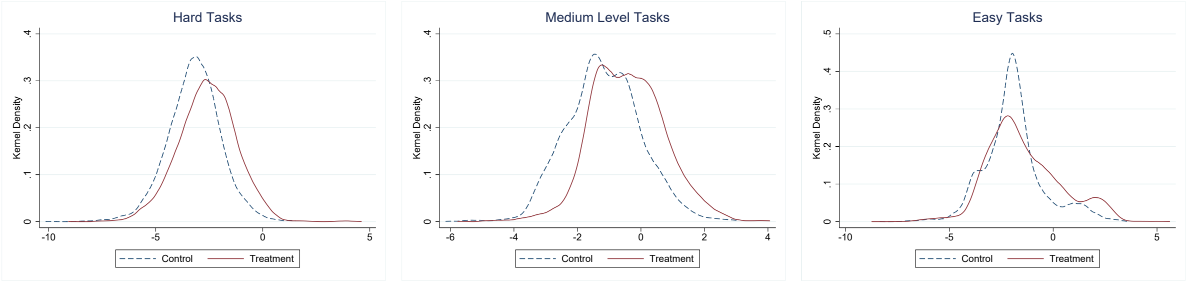
(b) Social-emotional Score

(c) Gross Motor Score

(d) Fine Motor Score

Figure 6: The Distribution of Latent Skill Loadings

Figures 6(a)-6(d) help to explain these results. The latent skill factor loadings play an important role. Figure 6 plots the estimated skill factor loadings $\boldsymbol{\alpha}_j$ for language and



(a) Hard Tasks (b) Median Level Tasks (c) Easy Tasks

Figure 7: The Distribution of $\boldsymbol{\alpha}_j'\boldsymbol{\theta}_i$

cognitive task items.[29] The loadings for the treatment group are larger for the harder tasks, while the size of loadings is larger for harder and easier tasks but smaller for tasks in the medium difficulty range. 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 $\alpha^{j,1} = \alpha^{j,0}$.[30]

Table 9: Treatment Effects on 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 10: 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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[29] Appendix H presents the latent skill loadings on other types of tasks.

[30] In Appendix J, Tables J3-J4 provide the tables for item-by-item tests. Social-emotional item loadings are not precisely estimated.

Table 11: Latent Skill Loadings on Denver Test Tasks

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

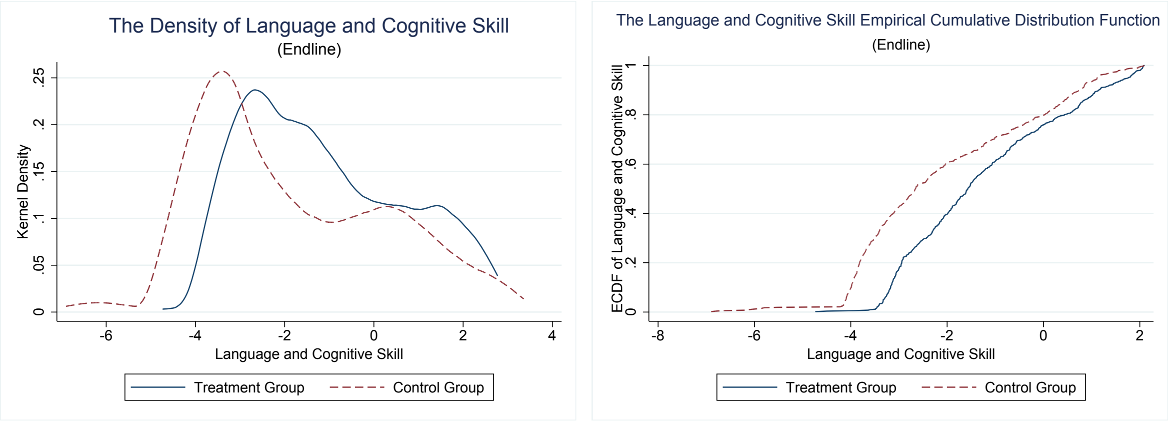
Notes: 1. These are the means and variances of $\alpha^{j,0}$ and $\alpha^{j,1}$, respectively, across items.

2. p-values are with respect to the null of equality of treatment and control summary measures.

As evident from equation (5), at the same level of skill, the larger the factor loadings, the better the child's performance. Table 11 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 of various tasks. For example, the loadings of language and cognitive skills are very large for language and cognitive tasks, but the loadings of social-emotional skills for the same tasks are relatively small. This means that, given the same amount of increase in language and social-emotional skills, language skills improve the language task performance more.

3.3.3 Distributions of Latent Skill

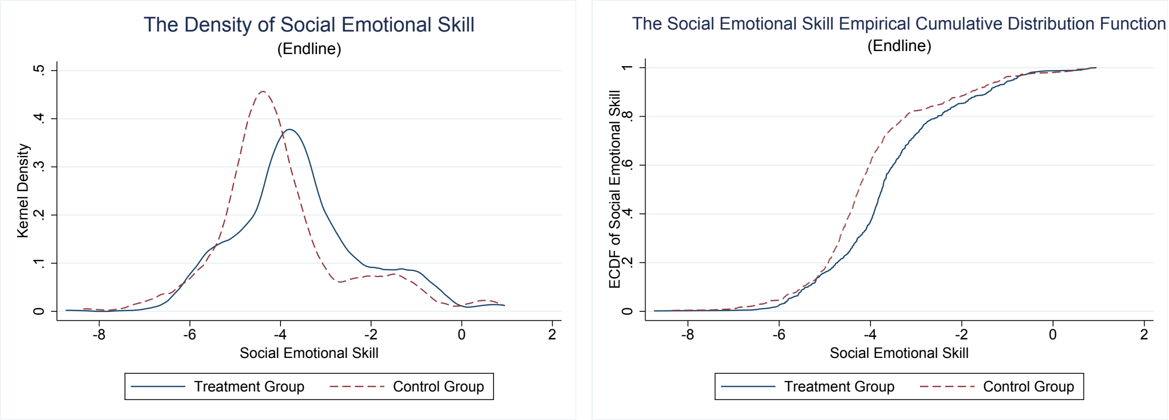
We first compare the language skill distributions of the control and treatment groups. Figure 8 (a) shows that the density of language and cognitive skills for the treatment group shift right; the treatment group shifts right and also has a fatter upper tail than the one in the control group. Figure 8 (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)

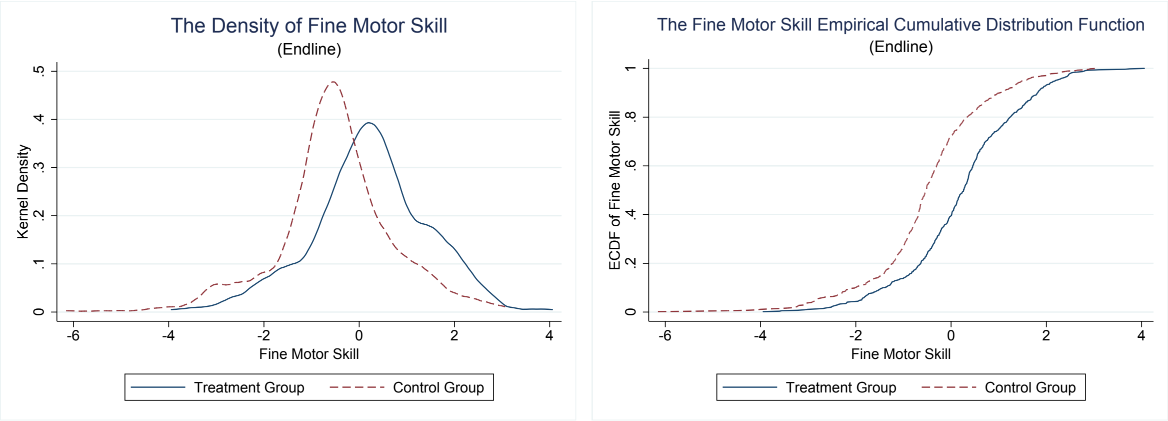
Figure 8: Language and Cognitive Skills Distribution and Stochastic Dominance Curves

Switching focus to social-emotional and fine motor skills, children in the treatment group are more concentrated at the upper level of the distribution, which is consistent with Figures 9 (a) and 10 (a) which present the probability density distribution of social-emotional and fine motor skills.



(a) (b)

Figure 9: Social-emotional Skills Distribution and Stochastic Dominance Curves



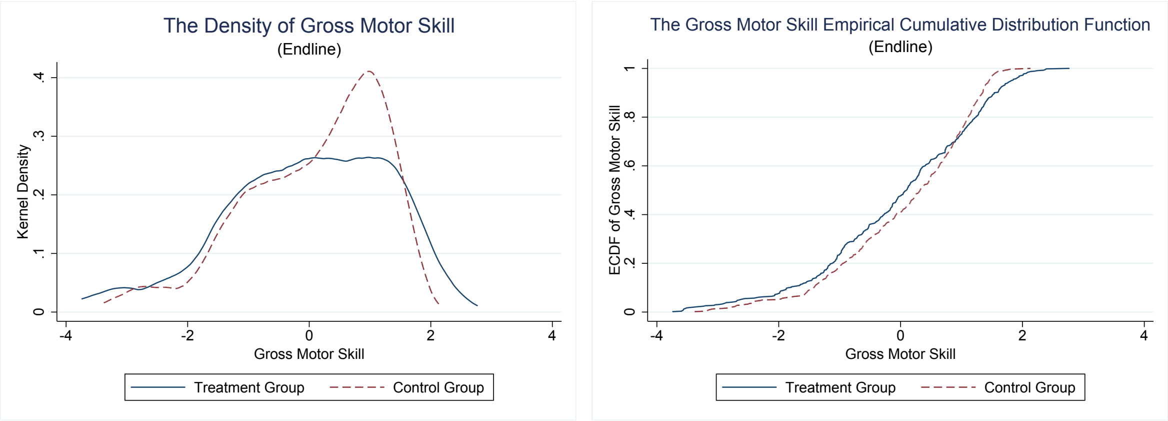
(a) (b)

Figure 10: Fine Motor Skills Distribution and Stochastic Dominance Curves

Regarding gross motor skills, we find that the factor distributions are similar between the control and the treatment groups. Figures 11 (a) and (b) show that both the densities and CDFs of the two gross motor skills distributions are close.

There are two main findings reported in this section. First, 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. The reason is that mean treatment effects show the combination effects of latent skills and the impact of the skill loadings. However, as we show below, the latter plays a minor role.

Second, gross motor distributions are not significantly different between the control and the treatment groups, which is also consistent with the mean treatment effect estimates. We next explore the sources of these treatment effects.



(a)

(b)

Figure 11: Gross Motor Skills Distribution and Stochastic Dominance Curves

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:

$$y_i^{j*} = \boldsymbol{X_i'}(D_i\boldsymbol{\beta^{j,1}} + (1-D_i)\boldsymbol{\beta^{j,0}}) + D_i((\boldsymbol{\theta_i^1})'\boldsymbol{\gamma^{j,1}}) + (1-D_i)((\boldsymbol{\theta_i^0})'\boldsymbol{\gamma^{j,0}}) + \varepsilon_i^j \tag{7}$$

where, 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 $\varepsilon_i^j \perp\!\!\!\perp \varepsilon_i^m, \forall\, m, j \in \{1, \ldots, N_J\}, m \neq j$

$$\underbrace{\sum_{j\in\{1,\ldots,N_J\}} y_i^{j*}}_{\text{Denver Test Score } \mathrm{Y}_i} = \sum_{j\in\{1,\ldots,N_J\}} \boldsymbol{x_i}(D_i\boldsymbol{\beta^{j,1}} + (1-D_i)\boldsymbol{\beta^{j,0}}) + D_i(\sum_{j\in\{1,\ldots,N_J\}} (\boldsymbol{\theta_i^1})'\boldsymbol{\alpha^{j,1}}) \tag{8}$$

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

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

$$\tilde{\lambda} := E(\sum_{j\in\{1,\ldots,N_J\}} (\boldsymbol{\theta_i^1})'\boldsymbol{\gamma}^{j,1}|\boldsymbol{x_i}, D_i = 1) - E(\sum_{j\in\{1,\ldots,N_J\}} (\boldsymbol{\theta_i^0})'\boldsymbol{\gamma^{j,0}}|\boldsymbol{x_i}, D_i = 0) \tag{9}$$

We ignore $\boldsymbol{X}$ because we cannot reject the hypothesis that $\boldsymbol{\beta^{j,1}} = \boldsymbol{\beta^{j,0}}$. Since we recover the individual latent skills $\boldsymbol{\theta_i^d}$, equation (9) 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 12.

The point estimates of the average treatment effects are almost identical using these two methods. From the column of p-values, we cannot reject the hypothesis that the two estimates are the same.

Table 12
Average Treatment Effect Point Estimates Comparison

Denver Tasks	Directly Obtained From Experiment	Derived From Latent Skill	p-value
	ATE	ATE	
Language and Cognitive	1.113 [0.723, 1.510]	1.115 [0.765, 1.454]	0.504
Social-emotional	-0.115 [-0.491, 0.275]	-0.081 [-0.315, 0.152]	0.556
Fine Motor	0.645 [0.139, 1.158]	0.569 [0.136, 0.990]	0.413
Gross Motor	0.219 [-0.294, 0.775]	0.190 [-0.071, 0.450]	0.460

Notes: 1. 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 Table 2 column (5).

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 treatment effects, and the effects of treatment on skill factors.

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

$$Y_i^{j,d} = 1(\boldsymbol{X_i'}\boldsymbol{\beta^{j,d}} + \delta^j + (\boldsymbol{\theta_i^d})'\boldsymbol{\alpha^{j,d}} + \varepsilon_i^j > 0) \tag{10}$$

where we assume $\varepsilon_i^j \sim N(0,1)$. From Equation (10), we see that the home visiting treatment effects come from three channels: 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:

$$\begin{aligned}
\Pr(Y^{j,1} = 1) - \Pr(Y^{j,0} = 1) \quad &= \quad \underbrace{\int \{\Phi([\boldsymbol{x'}\boldsymbol{\beta^{j,1}} + \delta^j + (\boldsymbol{\theta^1})'\boldsymbol{\alpha^{j,1}}]) - \Phi([\boldsymbol{x'}\boldsymbol{\beta^{j,0}} + \delta^j + (\boldsymbol{\theta^1})'\boldsymbol{\alpha^{j,1}}])\} dF^1(\boldsymbol{\theta^1}, \boldsymbol{X})}_{\text{From Estimated Coefficients of X}} \qquad (11) \\
&+ \quad \underbrace{\int \{\Phi([\boldsymbol{x'}\boldsymbol{\beta^{j,0}} + \delta^j + \boldsymbol{(\theta^1)'}\boldsymbol{\alpha^{j,1}}]) - \Phi([\boldsymbol{x'}\boldsymbol{\beta^{j,0}} + \delta^j + \boldsymbol{(\theta^1)'}\boldsymbol{\alpha^{j,0}}])\} dF^1(\boldsymbol{\theta^1}, \boldsymbol{X})}_{\text{From Latent Skill Loadings}} \\
&+ \quad \underbrace{\int \Phi([\boldsymbol{x'}\boldsymbol{\beta^{j,0}} + \delta^j + \boldsymbol{(\theta^1)'}\boldsymbol{\alpha^{j,0}}]) dF^1(\boldsymbol{\theta^1}, \boldsymbol{X}) - \int \Phi([\boldsymbol{x'}\boldsymbol{\beta^{j,0}} + \boldsymbol{(\theta^0)'}\boldsymbol{\alpha^{j,0}}]) dF^0(\boldsymbol{\theta^0}, \boldsymbol{X})}_{\text{From Latent Skill Factors}}.
\end{aligned}$$

Notice that equation (11) holds when there is common support for $\boldsymbol{X}$ and the factors in the control and treatment groups have similar distributions of observable covariates, which is essentially satisfied in our sample.[31] Table 13 reports the decomposition of treatment effects.[32] The main drivers of treatment effects are increases in latent skills. The

[31] 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.

[32] We set $\boldsymbol{\beta^{j,0}} = \boldsymbol{\beta^{j,1}}$ since it is consistent with the evidence.

contributions from experimentally-induced changes in $\boldsymbol{\alpha}$ are not precisely estimated.

Table 13: Sources of the Treatment Effects

Tasks	Total Net Treatment Effects	From Skill Loadings α	From Latent Skills θ
Language and Cognitive	1.096	0.126	0.970
	(0.312)	(0.135)	(0.174)
		11%	89%
Social-emotional	0.258	-0.034	0.292
	(0.131)	(0.084)	(0.078)
		-13%	113%
Fine Motor	0.303	-0.089	0.392
	0.164	(0.062)	(0.082)
		-30%	130%
Gross Motor	0.150	-0.078	0.228
	(0.153)	(0.072)	(0.093)
		-52%	152%

Notes: 1. The total treatment effects are defined as $T_k = \sum_{j \in k}(\sum_{i \in D^1} \mathbf{1^{j,1}} - \sum_{i \in D^0} \mathbf{1^{j,0}})$

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 month old.

3. Standard errors are reported in the parentheses.

Table 14: 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 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 two programs.

Table 14 shows that for comparable outcome measures at the early ages, 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.

5 Conclusion

This paper estimates the treatment effects from a large scale early childhood home visiting intervention program (China REACH) on child skill development, patterned after the successful and widely-emulated Jamaica Reach Up and Learn program. Since national policy in China is driven by evidence, rigorous analysis of 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 adjust 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, although the program shifts aggregate measures of the mappings from skills to tasks. Effects of the program appear to arise primarily from beneficial interaction patterns between home visitors and caregivers and home visitors and children, a point we develop further in a companion paper (Heckman and Zhou, 2020). Our analysis offers a prototype for measuring latent skills from 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. 3.3.1

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

Bai, Y., J. P. Romano, and A. M. Shaikh (2019). Inference in experiments with matched pairs. Unpublished manuscript, University of Chicago. 3.1

Cameron, C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *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. 13

Chen, M., I. Fernandez-Val, and M. Weidner (2019). Nonlinear factor models for network and panel data. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies. 3.3.1

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. 3.3.1

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. 3.3.1

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. 22

Gertler, P., J. J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. Chang, and S. M. Grantham-McGregor (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. *Science 344*(6187), 998–1001. 2, 4.2

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, 4.2

Heckman, J. J. and G. Karapakula (2019). The Perry Preschoolers at late midlife: A study in design-specific inference. NBER Working Paper 25888. 22

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 (2020). The impacts of child-caregiver and caregiver-home visitor interactions on child skill development. Unpublished. 2.1, 20, 5

Lu, B., R. Greevy, X. Xu, and C. Beck (2011). Optimal nonbipartite matching and its statistical applications. *American Statistics 65*(1), 21–30. 5

Ma, X. and J. Wang (2019). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*. 14

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. 10

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, 3.3.1