NBER WORKING PAPER SERIES

THE IMPACTS OF A PROTOTYPICAL HOME VISITING PROGRAM ON CHILD SKILLS

James J. Heckman
Bei Liu
Mai Lu
Jin Zhou

Working Paper 27356
http://www.nber.org/papers/w27356

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2020, Revised February 2022

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

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

The Impacts of a Prototypical Home Visiting Program on Child Skills
James J. Heckman, Bei Liu, Mai Lu, and Jin Zhou
NBER Working Paper No. 27356
June 2020, Revised February 2022
JEL No. J13,Z18

ABSTRACT

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

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

Mai Lu
China Development Research Foundation
Floor 15, Tower A
Imperial International Center
No. 136, Andingmen Wai Avenue
Dongcheng District
Beijing, P.C. 100011
China
lumai@cdrf.org.cn

Bei Liu
China Development Research Foundation
Floor 15, Tower A
Imperial International Center
No. 136, Andingmen Wai Avenue
Dongcheng District
Beijing, P.C. 100011
China
liubei@cdrf.org.cn

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

A data appendix is available at http://www.nber.org/data-appendix/w27356

1 Introduction

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

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

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

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

This paper proceeds as follows. Section 2 describes the program. It is a scaled and enhanced version of the original Jamaica program. Section 3 presents an array of conventional experimental treatment effects and documents heterogeneity in program impacts. Furthermore, we estimate a nonlinear factor model with

individual-level latent skills and determine the impact of treatment on the skills that generate item scores. Section 4 examines the sources of the estimated treatment effects. We examine the extent to which the program affects the inputs into the functions mapping skills to performance on tasks and the extent to which it shifts the productivity of a fixed stock of latent skills. Section 5 compares outcomes from the China program with those from the parent Jamaica program with follow-up through age 30. China REACH is on track to replicate Jamaica's long-term improvement of education and labor market outcomes. Section 6 summarizes our findings.

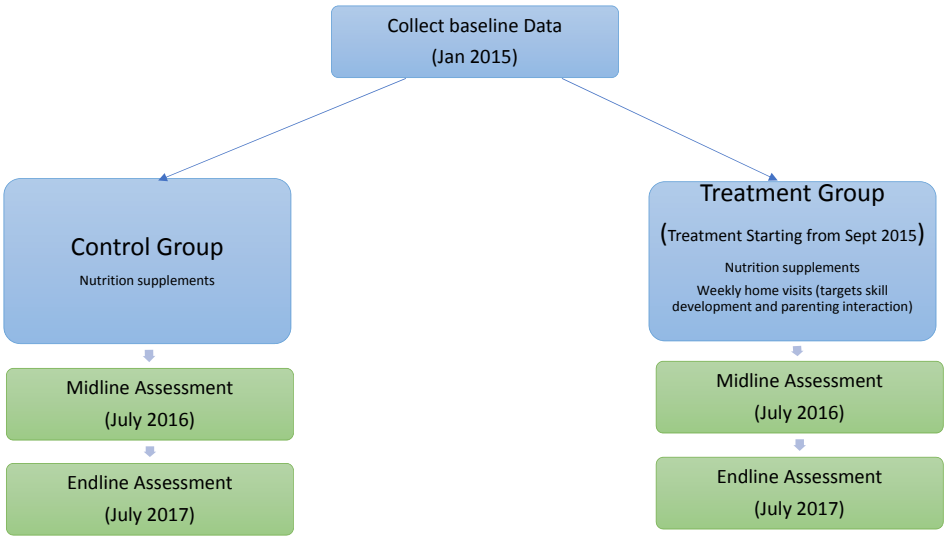
2 China REACH

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

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

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

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



2.1 The Intervention Implemented

The program trains home visitors who have educational attainments at the level of the mothers visited. In rural China, it is easily replicated because the potential supply of home visitors is large. The program encourages child caregivers to interact with their children in developmentally appropriate ways. Lizzeri and Siniscalchi (2008) develop a model of child development that features parent-child interactions as important determinants of good parenting.[2]

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 prepares countywide training to oversee the township supervisors. The county project coordinator and township supervisors randomly attend home visits for spot checks to observe and review the home visitors' work.

The supervisors support and manage home visitors. They make sure that the home visitors prepare for weekly visits, review the content of past visits, plan activities for future visits, and organize weekly meetings with the home visitors to

[2] Heckman and Zhou (2021) document the home visiting protocols used.

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

improve and reflect on the home visiting program and experience. Township supervisors visit each household with the home visitor once a month and record observations on the caregiver, child, and home visitor and their interactions.

The visitors engage with households weekly and provide one hour of parenting or caregiving guidance and support based on the Jamaica program protocols.[4] During each home visit, the home visitor records information about parental engagement (e.g., who worked with the child during the visit, whether the home visitor taught parents relevant tasks if the child could not participate in the home visit, and who played with the child after the visit and with what frequency) and child performance (e.g., tasks taught in the last week and new tasks taught in the current week). 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

Randomization is based on a village- (cluster-) level matched-pair design. Bai (2019) shows that this design is optimal for minimizing the mean-squared error of estimates of average treatment effects. Implementation is in three steps. We first examine the entire universe of eligible villages in Huachi county. We next use household surveys and village-level administrative data to assess the similarities of villages using a Mahalanobis metric of resident and village characteristics.[5] To minimize the Mahalanobis metric in each pair, we sort the villages by metric

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

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

scores and pair the closest ones 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 A.2 in the Appendix indicates the location of the paired villages in Huachi county. The design closely matches the characteristics of the villages in the pairs.[8] Village-level treatment effects include within-village spillovers. Villages are used only once, as either treatments or controls.

3 Estimated Treatment Effects

The China REACH intervention aims to promote multiple skills (e.g., motor, language, cognitive, and social-emotional skills). Table 1 displays our measures of skill. The Denver II test provides detailed child development assessment task measures.[9,10,11]

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

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

[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 observations rather than a parental report. It is an inventory of 125 tasks, including four types of skill measures: personal-social (caring for personal needs and getting along with people), fine motor–adaptive (hand-eye coordination, manipulation of small objects, and problem-solving), language (hearing, understanding, and using language), and gross motor (sitting, walking, jumping, and overall large muscle movement).

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

Table 1: China REACH Home Visiting Program Skill Content

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

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

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

3.1 County-Level Average Treatment Effects

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

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

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

twice.

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

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

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

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

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

Treatment is at the village level. The idiosyncratic shock term ε_{iv}^m for child i can be arbitrarily correlated with $\varepsilon_{i'v}^m$ for any other child $i' \neq i$ in the same village v. Idiosyncratic shocks are assumed to be independent across villages; i.e., $\varepsilon_{iv}^m \perp\!\!\!\perp \varepsilon_{kv'}^m$ for $\forall i \in v$ and $\forall k \in v', v \neq v'$. Residual plots displayed in Appendix D 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 that the ratio of clusters to observations in the cluster converges to a constant. This is true if $\boldsymbol{\beta_1^m}$ is constant across people.

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

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

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

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

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

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

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

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

[17] Only 140 children took the Denver test at the baseline. We estimate the same model for the children with baseline information and do not find significant differences in Denver test scores between the control and treatment groups. The details about this balancing test are presented in Appendix 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. (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 number of item scores comparable to the number we have for the other categories.

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

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

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

[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 D.

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

Table 2: Treatment Effects on Standardized Denver Scores

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

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

Table 3: Treatment Effects on Standardized Denver Scores

(Female)

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

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

Table 4: Treatment Effects on Standardized Denver Scores

(Male)

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

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

Appendix G presents an analysis of the impacts on child skills of interactions between the home visitor and the caregiver, and the home visitor and the child, as well as variables capturing home visitor teaching ability.[24] The only strong pattern that emerges is that good caregiver–home visitor interactions promote language and cognitive skills.[25]

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

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

To address this issue, we take advantage of the multi-item nature of our data and estimate a nonlinear factor model with individual-level latent skills.[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 map between skills and item scores (i.e., whether treated children better utilize existing skills).

3.2.1 Items and Skills

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

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

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

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

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

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

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

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

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

Under suitable normalizations, we can identify the *individual*-level latent skill factors $\boldsymbol{\theta_i^d}$ and not just the distribution of the latent skill factors, as in traditional psychometric models (see, e.g., van der Linden, 2016). We assume that $\varepsilon_i^{j_k}$ is unit normal, independent of the other right-hand side variables. This data has a panel-like structure over items. It can be fit using a probit model with latent skills. We estimate the parameters of observed covariates, the latent factors, and the effects of latent skill factors on outcomes. From the analysis of Wang (2020), it can be shown that estimators of the parameters of the model, including individual abilities, are consistent and asymptotically unbiased when the number of observations (sample participants) $N_I \to \infty$ and $N_J \to \infty$ but $\frac{N_I}{N_J}$ converges to a constant.[27] These con-

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

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

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

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

We have four different latent skill factors in our model, corresponding to social-emotional, language and cognitive, fine motor, and gross motor skills in the Denver II test $k \in \{1, \ldots, 4\}$. To interpret the factors, we assume that performance on K of N_J tasks ($K \leq N_J$) depends only on one factor. This specializes what Cunha, Heckman, and Schennach (2010) call the "dedicated factor case" to apply to only the first four items of each measurement. We thus generalize their analysis by requiring that only a subset of tasks are dedicated for any measurement of skills. We normalize the factor loading matrix so that the first K rows form an $\boldsymbol{I}_{K,K}$ identity matrix. For the first $K = 4$ items of the measurements, we assume that they load on only one skill.[28] The remaining factor loading matrix for the vector of N_J outcomes is unrestricted. Dropping the d superscript to reduce notational clutter, we write the metric of loadings on the latent skills as $\boldsymbol{\alpha}'_{N_J \times K}$:

design.

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

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

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

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

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

We report sensitivity analyses of our estimates using a variety of plausible normalizations in Appendix H. We find that the estimates of $\boldsymbol{\alpha}^{j_k,d}$ reported in the text are stable under a variety of different normalizations.[29] Our results are quantitatively robust. We use the estimation procedure proposed by Chen, Fernández-Val, and Weidner (2021) to estimate panel probit models with multiple latent skill factors.[30] The asymptotic justification for this approach for estimating individual-specific factors and population factor loadings is based on Wang (2020).

3.2.2 Estimates

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

[29] In Appendix H, 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.

[30] Details regarding the method are presented in Appendix I.

males are substantially more negative for the treatment group. Figure 2 compares the distribution of the predicted combined language and cognitive task items from our model and the actual task items.[31] We also fit the data well with the other types of tasks.[32]

Table 6: Estimates of the Coefficients of the Observed Covariates

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

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

[31] We combine language and cognitive tasks into one category because of the paucity of cognitive test items in our Denver test.
[32] See Appendix J.

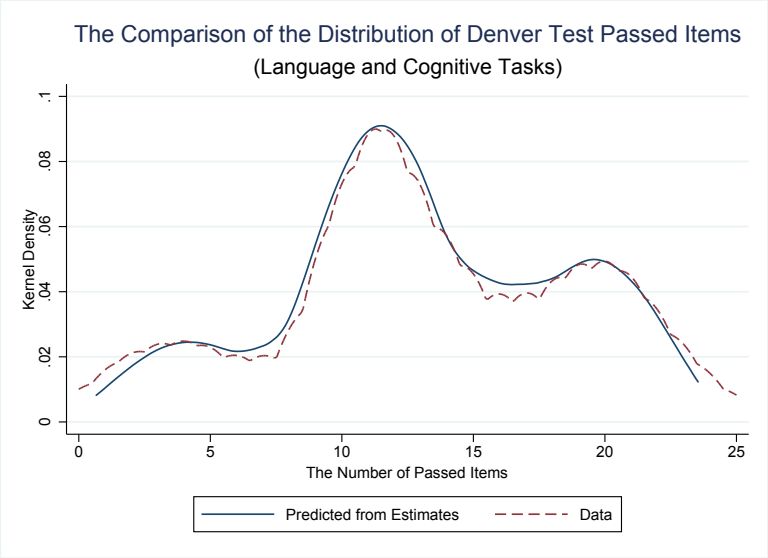


Figure 2: The Distribution of Denver Test Passed Items

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

Table 7: Treatment Effects on Mean of Latent Skill Factors

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

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

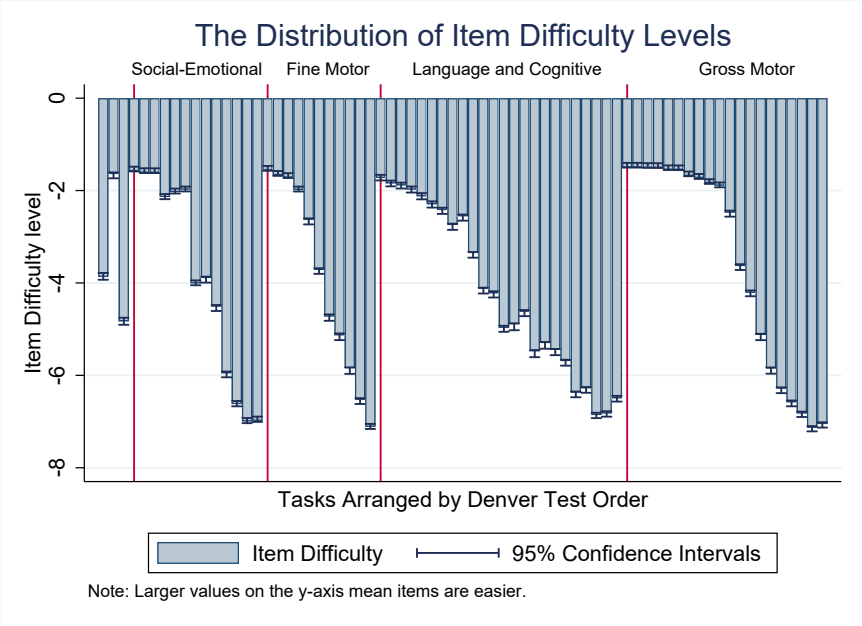


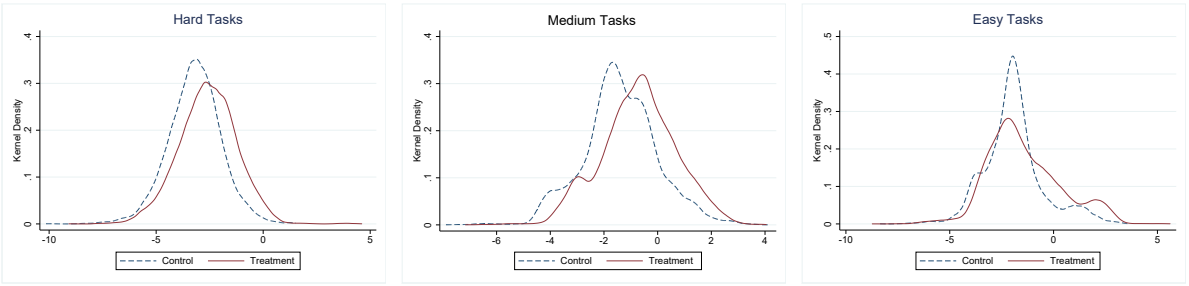
Figure 3: The Distribution of Denver Task Item Difficulty Levels

Table 8: The Correlation between Different Latent Skill Factors

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

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

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



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

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

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

Figure 4 plots the products of estimated skill factor loadings and the latent skill factors based on the Denver task difficulty levels.[33] The loadings for the treatment group are larger for the hard and medium tasks but smaller for the easy tasks, which indicates that the easier tasks are not helpful for detecting treatment effects on child skill development. The loadings have similar patterns across the treatment and control groups for other skills. Estimates of aggregates of loadings are precisely estimated, and for most tasks, we reject the hypothesis that $\alpha^{j_k,\ell,d=1} = \alpha^{j_k,\ell,d=0}$, $\ell \in \{1,\ldots,4\}$.[34] The only strong correlations are those between

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

[34] Tables H.3–H.4 in Appendix H provide item-by-item tests. Social-emotional item loadings are

social-emotional skills and fine motor skills.

Table 9: Skill Loadings on Denver Test Tasks ($\boldsymbol{\alpha}^{j_{k'}d}$) Latent Skills

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

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

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

3.2.3 Comparisons with a Model without Task Difficulty Parameters

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

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

not precisely estimated.

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

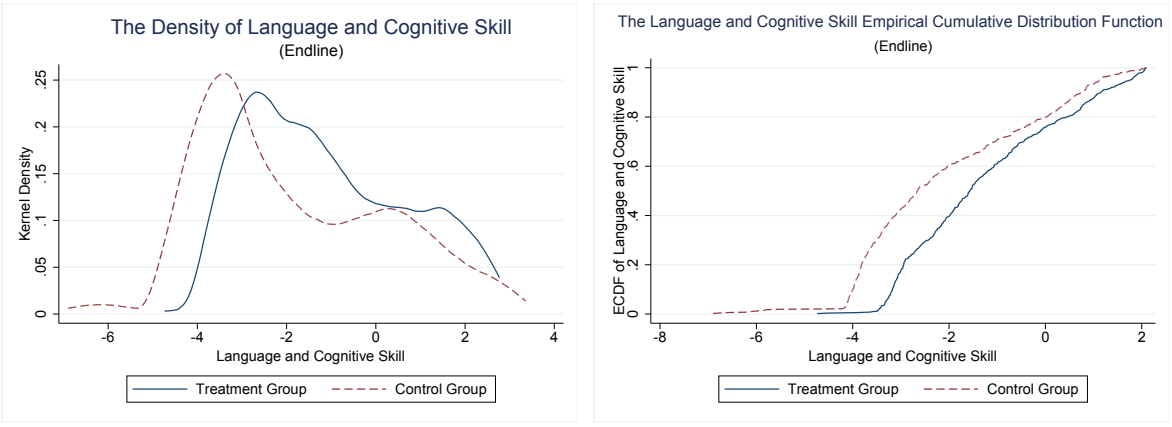
Table 10: Comparing Treatment Effects of θ_i Based on Two Models with and without Difficulty Parameters

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

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

3.2.4 Distributions of Latent Skill

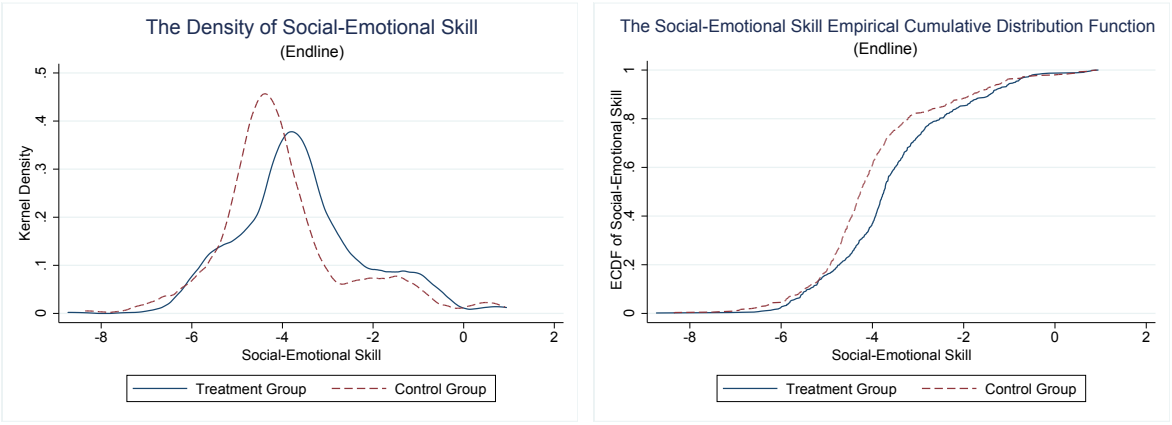
We compare the cognitive and language skill distributions of the control and treatment groups. Figure 5a shows that the density of language and cognitive skills for the treatment group shifts right and has a fatter upper tail than the one for the control group. Figure 5b shows that at almost every point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



(a) (b)

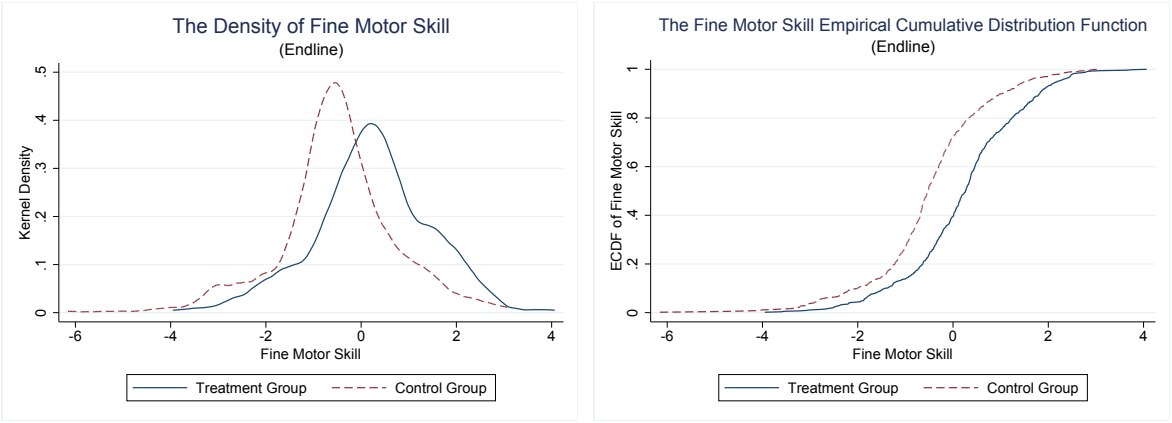
Figure 5: Language and Cognitive Skills Distribution

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



(a) (b)

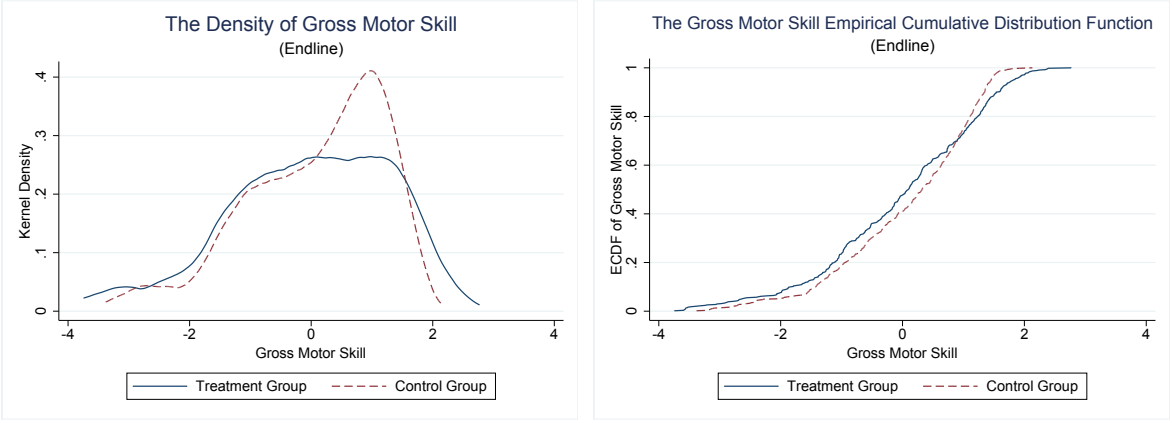
Figure 6: Social-Emotional Skills Distribution



(a) (b)

Figure 7: Fine Motor Skills Distribution

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



(a) (b)

Figure 8: Gross Motor Skills Distribution

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

4 Decomposing ATE

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

4.1 The Sources of Treatment Effects

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

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

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

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

Table 11: Average Treatment Effect Point Estimates Comparison

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

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

4.2 Decomposing Treatment Effects

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

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

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

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

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

Notice that equation (6) holds over a common support for $\boldsymbol{X}$ and when the factors in the control and treatment groups have similar distributions of observable covariates, which is essentially satisfied in our sample.[35] Table 12 reports the decomposition of treatment effects. The main drivers of the treatment effects are increases in latent skills. We have shown that there is no significant difference in $\boldsymbol{\beta}$ between the treatment and control groups in Table 6. Therefore, the contribution to the treatment effects from $\boldsymbol{\beta}$ is insignificant. The contribution from experimentally induced changes in $\boldsymbol{\alpha}$ is not precisely estimated. For this reason, we conclude that the dominant effect of treatment is on latent skills.

[35] 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 L, we show the age distribution between the treatment and control groups.

Table 12: Sources of the Treatment Effects

Tasks	Total Net Treatment Effects	From Observable Covariates	From Skill Loadings 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. Total treatment effects for skill k are $\frac{1}{N_{I_k}} \sum_{j_k=1}^{N_{I_k}} \left(\frac{\sum_{i=1}^{N_I} Y^{j_k,j} D_i}{\sum_{i=1}^{N_I} D_i} - \frac{\sum_{i=1}^{N_I} Y^{j_k,j}(1-D_i)}{\sum_{i=1}^{N_I}(1-D_i)} \right)$ assuming both denominators are nonzero and N_I is # of observations.

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

3. Standard errors are reported in parentheses.

4.3 Treatment Effects on Latent Skills Conditional on Caregiver Status

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

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

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

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

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

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

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

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

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

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

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

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

6 Conclusion

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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