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 January 2023

This paper was first presented at a session of the ASSA January 2020 Annual Meetings in San Diego organized by Scott Rozelle who attended the session along with Orazio Attanasio. We thank the editor, Kevin Lang, and three anonymous referees for their helpful comments. CEHD acknowledges support from the Institute for New Economic Thinking, the Eunice Kennedy Shriver National Institute of Child Health, and Human Development of the National Institutes of Health under award number R37HD065072, NICHD R01HD103666, and an anonymous donor. The program has been registered at AEA with registry number AEARCTR-0007119. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the funders or the official views of the National Institutes of Health. CDRF acknowledges support from the UBS Optimus Foundation and the Dunhe Foundation. The authors wish to thank Susan Chang, Sally Grantham-McGregor, Sylvi Kuperman, Carey Cheng, Rebecca Myerson, Chunni Zhang, and Yike Wang for their efforts on program design, implementation, and data cleaning support. Erlfang Tsai and Fuyao Wang provided highly competent research assistance. CDRF thanks Mary Young, Fan Bu, Peng Liu, Lijia Shi, Bojiao Liang, and Yi Qie for their essential and valuable fieldwork support. We are grateful to the participants and their families for their continued participation in this research project. http://cehd.uchicago.edu/china-reach_home-visiting_appendix is a website for this paper with supplementary material. Corresponding author: Jin Zhou (jinzhou@uchicago.edu

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 January 2023
JEL No. J13,Z18

ABSTRACT

This paper develops a new framework for estimating the causal impacts on child skills and the mechanisms producing these impacts using data from a randomized control study of a widely evaluated early-childhood home visiting program. We show the feasibility of replicating the program at scale. We report estimates from standard procedures for reporting treatment effects as unweighted averages item scores and compare them with estimates adjusting for item difficulties. Such adjustments produce more interpretable estimates. We go beyond treatment effects and estimate individual-specific latent skills, comparing treatment and control skills and their impacts on 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

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

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

Jin Zhou
Center for the Economics of
Human Development
University of Chicago
1126 East 59th Street
Chicago, IL 60637
jzhou269@cityu.edu.hk

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 for developing the skills of disadvantaged children.[1] They are relatively low cost compared to many other early childhood interventions. They place minimal demands on the training required of the visitors and on the infrastructure needed to support them.

This paper studies an implementation of the Jamaica Reach Up and Learn program at scale in China. Jamaican program, established over 30 years ago, is a successful, widely-emulated home visiting program. Visitors have levels of education comparable to those of the caregivers visited. This feature facilitates scalability. Its low cost and flexible format make it an appealing program for developing countries.[2]

This paper studies impacts on child skills and parental engagement of a large-scale replication of the original Jamaica program in a poor region of Western China. China Rural Education and Child Health program (China REACH) has 1500+ participants compared to the 100+ participants in the original Jamaica study and the typical size of other attempted replications. The program we study was designed by the architects of the original Jamaica program.

While the main features of the curricula of the two programs are identical, intake is not. Jamaica targeted stunted children. Its Chinese counterpart enrolls all children in a poor "rural" region of the country, save for the most physically compromised. Both programs are evaluated by a randomized control trial. Our evidence to date suggests that the program can be successfully implemented at

[1] See, e.g., HomVEE (2020) and Howard and Brooks-Gunn (2009).

[2] See Grantham-McGregor and Smith (2016) and Jervis et al. (2023) for a comprehensive list of programs implementing versions of the Jamaican program in different settings.

scale and that China REACH is on track to replicate Jamaica Reach.

The program improves home environments and multiple skills. We estimate *each child's* latent skills instead of just distributions of latent skills as is customary (see, e.g., Cunha and Heckman, 2008; Cunha et al., 2010 and in previous attempted replications). The intervention has a strong impact on language and cognitive skills, fine motor skills, and social-emotional skills, but impacts are not uniform across baseline skill distributions. Positive impacts are strongest for children with absent mothers. The China REACH program has much richer data than the original Jamaica program, in part because the same group of scholars designed both projects and incorporated lessons learned from Jamaica into the China replication.

We depart from conventional practice that measures achievement by the fraction of correct answers on items of tests and instead adjust for mastery of tasks by their difficulty. Adjusting produces more plausible estimated treatment effects. We go beyond conventional practice in the literature which assumes that only a single skill affects a particular measure of skill (called the "dedicated" case in the literature, e.g., Cunha et al., 2010). Multiple latent skills generate test scores instead of a single skills as assumed in most applications. Following Heckman et al. (2013), we decompose estimated treatment effects into enhancements in latent skills and improvements in the ability to use existing skills.[3] We find little evidence for the latter. Treatment effects primarily arise from boosts in skills.

This paper proceeds as follows: Section 2 describes the program we study. It is a scaled version of the original Jamaica program. Section 3 presents conventional experimental treatment effects and documents heterogeneity in program impacts and beneficial effects of the program on home environments. We also estimate multiskill factor models and construct individual-level latent skills. We determine

[3] This distinction goes back to Welch (1970).

the impact of treatment on the skills that generate responses to item scores. Section 4 examines the sources of the estimated treatment effects. 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 as documented in Gertler et al. (2014). Section 6 places the results of this paper in the context of a large literature that attempts to replicate the Jamaica program in a variety of different contexts. Section 7 summarizes our findings.

We exploit additional data from the program in our other research (Heckman and Zhou, 2022a,b,c). That research uses weekly data on the growth of skills only for *treatment group* children to understand the dynamics of skill formation. That research cannot measure treatment effects and is quite distinct from the analysis of program treatment effects presented in this paper and from previous analyses of the Jamaica program.

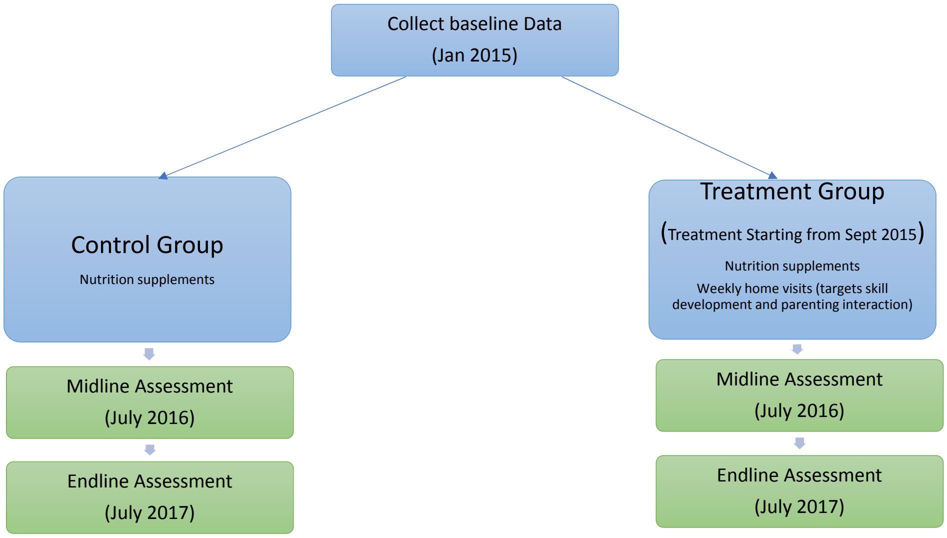
2 China REACH

The China Rural Education and Child Health (China REACH) project was launched in 2015 in response to a growing focus on, and call for, evidence-based pilot-to-policy analyses by China's State Council. It is a large-scale randomized control trial (RCT) designed to evaluate the impacts of a low-cost home visit delivery model for disadvantaged families. It is based on the curriculum of a successful Jamaican pilot.[4] The program aims to improve the health and cognition of children by enhancing their engagement with caregivers with program-enhanced skills.

[4] See Grantham-McGregor and Smith (2016) and Gertler et al. (2014), Gertler et al. (2022), and Jervis et al. (2023).

The program was conducted in Huachi County in Gansu Province, one of the poorest areas in China. The county has 15 townships in 111 administrative villages. For analytical convenience, two closely adjacent villages[5] are combined so there are 110 villages analyzed in this study. Huachi is 85% mountainous with a population of 132,000, of whom 114,600 have rural hukou.[6]

Figure 1: Timeline of China REACH (Huachi) Program



The baseline data collection was launched in January 2015 and home visits started in September 2015 (see Figure 1). We collected baseline data for all villages in Huachi county which covers the household economic demographic information and the measures of home environment (i.e., Infant/Toddler HOME Inventory scores). Given the baseline information, we designed our randomized control trial (see Section 2.1.1). We collect midline (about 9 months after the intervention) and endline (about 21 months after the intervention) data. At both midline and

[5] Chenghao and Wujiao.

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

endline, we collect the information for both control and treated groups including HOME score measures and relevant economics and demographic measures at the household level. For details on program implementation, see Appendix A.

2.1 The Intervention

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. Appendix B documents the home visiting protocols used.

Local implementation of the China REACH project is conducted by a county project coordinator, assisted by 24 township supervisors and 91 home visitors.[7] The coordinator prepares countywide training to oversee the township supervisors. The county project coordinator and township supervisors randomly attend home visits for spot checks to observe and review the home visitors' work. The supervisors have three years more education than that of the visitors, whose level of education is, on average, at the level of the mothers visited.

Supervisors support and manage the home visitors. They ensure that home visitors prepare for weekly visits, review the content of past visits, plan activities for future visits, and organize weekly meetings with the home visitors to improve and reflect on the home visiting program and experience. Township supervisors visit each household with the home visitor at least once a month and record observations on the caregiver, child, and home visitor and their interactions.

Home visitors engage with households weekly and provide one hour of parent-

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

ing or caregiving guidance and support based on the Jamaica program protocols.[8] The intervention does not target children directly, but instead targets caregivers (primarily mothers) who spend the most time with their children. During each home visit, the home visitor records information about parental engagement (e.g., who worked with the child during the visit, whether the home visitor taught parents relevant tasks if the child could not participate in the home visit, and who played with the child after the visit and with what frequency) and child performance (e.g., tasks taught in the last week and new tasks taught in the current week). Appendix B.3 documents the protocol of the China REACH program, 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, about 70 fine motor tasks, and 20 tasks targeting gross motor skill development. Denver tests are given at midline and endline. Denver II test are evaluated at midline and endline for both control and treatement children.

2.1.1 Design of the Randomized Control Trial

Randomization is based on a village (cluster) level matched-pair design. Bai (2022) shows that this design is optimal for minimizing the mean-squared error of estimates of average treatment effects. The experimental design guarantees exogeneity of regressors and identifies the parameters of the underlying models that generate the estimates.

Implementation is in three steps. First, the entire universe of eligible villages in Huachi county is examined. We use household surveys and village-level ad-

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

ministrative data to assess the similarities of villages using a Mahalanobis metric based on resident and village characteristics.[9] For our sample of 110 villages we form $\frac{1}{2}\left[110 \times (110-1)\right]$ metrics between pairs of villages. The second step generates 55 pairs and minimizes the sum of Mahalanobis distance of all pairs. Villages are sorted by their Mahalanobis score and pair the closest ones using the nonparametric belief propagation (nbp) matching method.[10] The nbp matching method constructs the pairs to minimize the sum of the Mahalanobis distances of 55 pairs. Bai (2022) shows that use of the Mahalanobis matrix has better performance than other metrics in generating smaller mean-squared errors for average treatment effects.

The third step randomly selects one village within each pair into the treatment group and the other paired village into the control group.[11] Figure A.2 displays the location of the paired villages in Huachi county. The design closely matches the characteristics of the villages in the pairs.[12] Village-level treatment effects include within-village spillovers. Villages are used only once, as treatments or controls.

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

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

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

[12] Appendix C documents baseline comparisons.

3 Estimated Treatment Effects

China REACH aims to promote multiple skills (e.g., motor, language, cognitive, and social-emotional skills). Table 1 displays our measures of skill. The Denver II test gives a detailed assessment of child development.[13],[14]

Table 1: China REACH Home Visiting Program Skill Content

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

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

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

[14] The Bayley III test converts composite scores into scaled scores based on age, which are more often used 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.

(Cameron et al., 2008).

Using proportions of items correctly answered standardized by sampling standard deviations (effect size) as outcomes is the standard practice in much of the evaluation literature and in all previous evaluations of attempted replications of Jamaica. This practice assumes that 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. In Section 3.3, we address this problem using a measurement model that accounts for variations in item difficulty[15] and recovers *individual* latent skills that generate item responses.

3.1 Estimating Average Treatment Effects

Following Bai et al. (2021) and Bai (2022), we report the treatment effects for a paired matching design. Our notation is as follows: 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:

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

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

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

[15] See, e.g., van der Linden (2016) for an exposition of this model.

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

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

where $\boldsymbol{Y_{iv}^{m}}$ 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. $\boldsymbol{Y_{iv}^{m}} = D_{v(i)}\boldsymbol{Y_{iv}^{m}}(1) + (1 - D_{v(i)})\boldsymbol{Y_{iv}^{m}}(0)$, where $\boldsymbol{Y_{iv}^{m}}(d)$ denotes the vector of outcomes fixing treatment status d. The randomized design implies that

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

The idiosyncratic shock term $\boldsymbol{\varepsilon_{iv}^{m}}$ for child i can be arbitrarily correlated with $\varepsilon_{i'v}^{m}$ for any other child $i' \neq i$ in the same village v. However, idiosyncratic shocks are assumed to be independent across villages; i.e., $\boldsymbol{\varepsilon_{iv}^{m}} \perp\!\!\!\perp \boldsymbol{\varepsilon_{kv'}^{m}}$ for $\forall i \in v$ and $\forall k \in v', v \neq v'$. Residual plots displayed in Appendix E validate the assumption. The $N \times N$ covariance matrix $E(\boldsymbol{\varepsilon\varepsilon'}) = \boldsymbol{\Omega}$ with V number of villages is block diagonal: $\boldsymbol{\Omega}_{vv'} = 0$; all $v \neq v'$.[16]

Define the full array of right-hand side variables in Equation (1) by $\boldsymbol{X_{iv}}$. The standard cluster-robust variance estimator (CRVE), $(\boldsymbol{X'X})^{-1}(\sum_{v=1}^{V}\boldsymbol{X_v'\hat{\Omega}_v X_v})(\boldsymbol{X'X})^{-1}$, is biased when $\hat{\Omega}_v$ is estimated using the OLS residuals $\hat{\varepsilon}_v$: $E(\boldsymbol{\hat{\varepsilon}_v\hat{\varepsilon}_v'})$. The bias depends on the form of $\boldsymbol{\Omega_v}$. Cameron et al. (2008) discuss this problem and show that the wild cluster bootstrap performs well in making cluster-robust inferences. Details of the wild bootstrap procedures used are presented in Appendix F.[17]

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

[17] Because we have 55 clusters, recent concerns about the wild bootstrap do not apply. See Canay

In our sample, over 98% of eligible children in the treated villages receive home visits. Still, about 15% of children from both treatment and control groups miss the annual child development assessment. In an effort to obtain consistent estimates of population average treatment effects, we use inverse probability weighting (Tsiatis, 2006).[18,19] In Table 2, we report estimates with IPW and in Appendix H, we show results without adjusting. In estimating our latent factor model, we also weight the observations. Treatment effects are for the village level and include any population spillovers.

Table 2 presents overall standard treatment effects (effect sizes) for each skill category using standardized outcome measures.[20,21] Columns (1), (2), and (3) are estimates based on all available data, and columns (4) - (6) only use samples of children who were under 2 years of age in September 2015 when the program started. The younger treated children have at least one year of exposure to the intervention.[22] Our estimates are robust whether or not we use matching instead of IPW-weighted OLS. See Appendix I.

et al. (2021).

[18] 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.

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

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

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

[22] 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.

The first row in Table 2 shows that children in the treatment group are, on average, more likely to have higher language and cognitive skills.[23] The first row shows that at midline (about nine months into the intervention) the language and cognitive skills of the children in the treatment group are about 0.7 standard deviations higher than those of children in the control group. At the end of the intervention, effect sizes for treatment effects on language and cognitive skills are greater than 1. The intervention significantly improves treated children's language and cognitive skills. Column (2) restricts the sample to children who were less than 2 years old at enrollment. Doing so generates a more age-balanced sample between treatment and control groups. In Appendix Figure J.1, we show that the monthly age distributions are comparable between treatment and control groups.

The intervention significantly improves social-emotional skills at midline and fine motor skills at the end of the intervention but produces no significant improvement in gross motor skills. This finding is consistent with the design of the curriculum, which focuses primarily on language and cognitive skill development.[24] See Appendix B.

The remaining columns of Table 2 display treatment effects by gender. An interesting finding, consistent with recurrent findings in the literature (Elango et al., 2016), is that the intervention improves boys' language and cognitive skills much more than those of girls. At midline, the treatment effect sizes are 0.4 for girls and 0.9 for boys. At the end of the intervention, the effect sizes are about 0.9 for girls and boys. Non-intuitively, socio-emotional skills are reduced at endline. This result vanishes when we account for item difficulty.

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

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

Table 2: Treatment Effects on Standardized Denver Scores

	(1)	(2)	(3)	(4)	(5)	(6)
	All Enrolled Children			Children $\leq$ 2 Yrs at Enrollment		
	Both Gender	Female	Male	Both Gender	Female	Male
Midline						
Language and Cognitive	0.714***	0.445	0.938***	0.741***	0.534**	0.911***
	[0.319, 1.093]	[-0.014, 0.910]	[0.389, 1.499]	[0.350, 1.144]	[0.080, 0.990]	[0.329, 1.501]
Fine Motor	0.633*	0.335	0.716	0.703*	0.544	0.771
	[0.003, 1.313]	[-0.269, 1.211]	[-0.099, 1.598]	[0.057, 1.375]	[-0.082, 1.189]	[-0.070, 1.747]
Social-Emotional	0.879***	1.114***	0.549**	0.620***	0.938***	0.280
	[0.467, 1.289]	[0.681, 1.550]	[0.047, 1.054]	[0.204, 1.067]	[0.400, 1.431]	[-0.272, 0.842]
Gross Motor	-0.015	0.058	-0.041	0.010	0.019	-0.021
	[-0.567, 0.554]	[-0.532, 0.675]	[-0.700, 0.639]	[-0.559, 0.584]	[-0.605, 0.652]	[-0.682, 0.659]
Endline						
Language and Cognitive	1.036***	0.950**	0.950***	1.113***	0.893**	1.111***
	[0.644, 1.458]	[0.213, 1.675]	[0.448, 1.497]	[0.723, 1.510]	[0.177, 1.598]	[0.625, 1.626]
Fine Motor	0.676***	0.866**	0.462	0.645**	0.855**	0.388
	[0.180, 1.170]	[0.189, 1.574]	[-0.206, 1.144]	[0.139, 1.158]	[0.117, 1.579]	[-0.355, 1.124]
Social-Emotional	-0.222	-0.309	-0.256	-0.115	-0.291	-0.169
	[-0.636, 0.194]	[-0.775, 0.160]	[-0.829, 0.326]	[-0.491, 0.275]	[-0.820, 0.206]	[-0.701, 0.400]
Gross Motor	0.173	0.257	-0.048	0.219	0.445	-0.138
	[-0.322, 0.668]	[-0.582, 1.080]	[-0.510, 0.419]	[-0.294, 0.775]	[-0.417, 1.326]	[-0.629, 0.359]
Pre-Treatment Covariates	Yes	Yes	Yes	Yes	Yes	Yes
IPW	Yes	Yes	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. 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. 4. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
5. The negative treatment effects for social-emotional ability vanish after we adjust for item difficulty.

Appendix H presents a version of Table 2 that reports results for alternative statistical procedures: (a) use IPW or not, and (b) adjust for pretreatment variables or not. The qualitative results are similar to the results in Table 2 and application of none of the alternative procedures alters our account of Table 2. Our estimates are robust when we use matching instead of IPW-weighted OLS. See Appendix I.

3.2 Impacts on Home Environments

The program was designed to improve the home lives of treated children. Data are collected for both treatment and control groups on home environments as assessed by supervisors. Table 3 reports the treatment effects on HOME environment scores. The intervention significantly improves the composite HOME scores.[25]

Table 3: Treatment Effects on Home Environment Scores

	(1)	(2)
	All	Children $\leq$ 2 Yrs at Enrollment
Home Total	0.868***	0.720**
	[0.309, 1.409]	[0.159, 1.269]
Home Involvement	0.241***	0.201***
	[0.109, 0.367]	[0.073, 0.327]
Home Variety	0.114	0.093
	[-0.025, 0.253]	[-0.037, 0.224]
Home Responsivity	0.066	0.048
	[-0.169, 0.300]	[-0.192, 0.289]
Home Acceptance	0.059	0.044
	[-0.041, 0.157]	[-0.064, 0.150]
Home Organization	0.095	0.069
	[-0.059, 0.242]	[-0.077, 0.223]
Home Learning Materials	0.291*	0.262*
	[0.047, 0.533]	[0.007, 0.512]
Pre-Treatment Covariates	Yes	Yes
IPW	Yes	Yes

Notes: 1. The 90% confidence intervals in brackets are constructed using the wild bootstrap clustered at the village level. 2. The column with the label "All" includes all the observations, and the column with the label "Children $\leq$ 2 Yrs at Enrollment" restricts the sample to the children who were under 2 years old when they enrolled in the program. 3. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

[25] The home involvement category is based on the definition in the Infant-Toddler HOME Inventory: Involvement including the following items: (1)Parent keeps child in visual range, looks at often. (2)Parent talks to child while doing household work. (3) Parent consciously encourages developmental advance. (4) Parent invites maturing toys with value via personal attention (5) Parent structures child's play periods, and (6) Parent provides toys that challenge child to develop new skills.

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

The preceding analysis follows standard practice and reports unweighted item aggregates. Such aggregates, while traditional, are problematic unless the difficulty is the same across all items, which is not generally 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.[27] We also estimate individual-level latent skills rather than distributions of skills as in traditional. We use our estimates to determine the impact of treatment on the skills that generate item scores. Following Heckman et al. (2013), we also estimate whether treated children better utilize existing skills than untreated children.

3.3.1 Items and Skills

We study children's performances on individual items of tests. There are N_{J_k} tasks (items) for each of the K distinct skills. Tasks are skill-specific (e.g., motor, cognitive, language, etc). Performance on the tasks is assumed to be generated by a *vector* of latent skills $\boldsymbol{\theta}$. Unlike most work in psychology and economics, we do not use a "dedicated" factor model that assumes that only one component of $\boldsymbol{\theta}$ affects test outcomes. Let N_J denote the total number of items for all skills (i.e., $N_J = \sum_{k=1}^{K} N_{J_k}$). Assume that a common technology maps skills to test scores in

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

[27] See von Davier (2016) for discussion of item difficulty parameters in the Rasch model.

all villages, dropping 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, $i \in \{1, \ldots, I\}$. 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 a person with treatment status d. $\boldsymbol{X_i}$ is a vector of baseline covariates. Individual skill components are assumed to be independent ($\theta_m \perp\!\!\!\perp \theta_n$). Write the mapping from latent skills $\boldsymbol{\theta_i^d}$ to the determinants of outcome on task j as

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

Define

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

where $\boldsymbol{\alpha^{j_k,d}}$ is a K-dimensional vector of factor loadings; δ^{j_k} is an item difficulty parameter for the item j_k; and the coefficients $\boldsymbol{\beta^{j_k,d}}$ and $\boldsymbol{\alpha^{j_k,d}}$ can depend on treatment, the particular skills modeled, and even the item studied, where items are common across people. In estimation, we impose $\boldsymbol{\beta^{j_k,d}} = \boldsymbol{\beta^{j_k',d}}$, $\forall j_k$ and j_k'; i.e., coefficients are common across all items. Specification (3) generalizes the standard scalar item response model of psychometrics to incorporate a vector of skills.[28]

This model interprets the intervention as shaping skills that affect performance on tasks (items). The intervention may also enhance the productivity of any given

[28] van der Linden (2016) discusses the single-skill item response model and also discusses versions of it with vectors of skill. See Rabe-Hesketh and Skrondal (2016) for a vector version. Carneiro et al. (2003) develop the model we use and estimate it by MCMC. See also Cameron and Heckman (1998, 2001); Heckman (1981); Muthen (1984)

skill in performing a task; i.e., the intervention may shifts $\boldsymbol{\alpha}^{j_k,d}$. The expression $(\boldsymbol{\theta}_i^d)'\boldsymbol{\alpha}^{j_k,d}$ is a bundle of effective skills for outcome j_k from intervention $D = d$ arising from either source.

Under suitable normalizations, we can identify the *individual*-level latent skill factors $\boldsymbol{\theta}_i^d$ and not just the distribution of the latent skill factors, as in traditional psychometric models (see, e.g., van der Linden, 2016). We assume that $\varepsilon_i^{j_k}$ is unit normal, independent of the other right-hand side variables. Our data has a panel-like structure across 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. Wang (2020) shows that estimators of the parameters of the model, including individual abilities, are consistent and asymptotically unbiased when the number of observations (sample participants N_I) and the number of items (N_J) become large, $N_I \to \infty$ and $N_J \to \infty$ but $\frac{N_I}{N_J}$ converges to a constant.[29] These conditions apply in our sample with large numbers of test items per person (≥ 70 for each skill) and some 1500 observations.

If one seeks to isolate $\boldsymbol{\theta}^d$ from $\boldsymbol{\alpha}^{j_k,d}$, factor models require normalizations since $(\boldsymbol{\theta}_i^d)'\boldsymbol{\alpha}^{j_k,d} = [(\boldsymbol{\theta}_i^d)'\boldsymbol{A}][\boldsymbol{A}^{-1}\boldsymbol{\alpha}^{j_k,d}]$. 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, $(\boldsymbol{\theta}_i^d)'\boldsymbol{\alpha}^{j_k,d}$. We report such estimates. This answers the question of how treatment affects skill through whichever channel. However, following Heckman et al. (2013), it is also interesting to break out the impact of the intervention from each source.

We use a widely-used normalization originally suggested by Anderson and Rubin (1956) and separately identify both the vectors $\boldsymbol{\theta}_i^d$ and $\boldsymbol{\alpha}^{j_k,d}$.[30] Williams (2020)

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

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

suggests a variety of alternative normalizations which might instead fruitfully be used. This normalization enables us to examine the impacts of the intervention on endowments and the impacts of the intervention on the efficiency of agents in using skills. We report estimates for $\boldsymbol{\theta}_i^d$ 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 and more general item response 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 assumption facilitates comparability of measurements across treatments and controls. If the difficulty levels are different across treatments and controls, it is not possible to make meaningful comparisons across items.

We have four different latent skill factors in our model, corresponding to socio-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 the first K of the N_J tasks $(K \leq N_J)$ depends only on one factor. This generalizes what Cunha et al. (2010) and an entire literature call the "dedicated factor case" to apply to only the first four items of each measurement. Thus, instead of requiring that each row depends on only one factor, as in the "dedicated case", we only require that a *subset* of rows are dedicated to one factor for all measurement of skills. We place restrictions on the first rows of the factor loading matrix, the remaining factor loading matrix is unrestricted. Dropping the d superscript to reduce notational clutter and focusing on either the treatment of control group for simplicity, we write the metric of loadings on the latent skills as $\boldsymbol{\alpha}'_{N_J \times K}$:

$$\boldsymbol{\alpha}'_{N_J \times K} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \alpha^{2,1} & 0 & 0 & 0 \\ \alpha^{3,1} & 1 & 0 & 0 \\ \vdots & \alpha^{4,2} & 0 & 0 \\ \vdots & \cdots & 0 & 0 \\ \vdots & \cdots & 1 & 0 \\ \vdots & \cdots & \alpha^{6,3} & 0 \\ \vdots & \cdots & \cdots & 1 \\ \vdots & \cdots & \cdots & \alpha^{7,4} \\ \alpha^{N_J,1}, & \alpha^{N_J,2} & \alpha^{N_J,3} & \alpha^{N_J,4}. \end{bmatrix} \tag{4}$$

We test and reject the "dedicated model" that assumes that, for $j_k \geq 8$, $\alpha^{j_k,\ell,d} = 0$. Table 4 reports this test. The widely-used assumption of a dedicated factor model fails in our sample. For a proof of identification, see Carneiro et al. (2003) or Appendix K.

Table 4: Test of Hypothesis $\alpha^{j_k,\ell,d} = 0$ for $j_k \geq 8$

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

We report sensitivity analyses of our estimates using a variety of plausible normalizations in Appendix L. We find that the estimates of $\boldsymbol{\alpha}^{j_k,d}$ reported in the text are stable under a variety of different normalizations.[31] We use the estimation pro-

[31] In Appendix L, we compare the distribution of the skill loadings under different normaliza-

cedures proposed by Wang (2020) and Chen et al. (2021) to estimate panel probit models with multiple latent skill factors.[32]

3.3.2 Estimates

In Appendix N, Table N.1 presents estimates of $\boldsymbol{\beta}^{k,d}$. There are no statistically significant differences between the treatment and control groups, although the point estimates for males are substantially more negative for the treatment group. Figure 2 compares the distribution of the predicted combined language and cognitive task items from our model with difficulty parameter and the actual task items.[33] We fit the data as well when we investigate the other types of tasks.[34] We find qualitatively similar results when we use a richer set of covariates. See Appendix Table P.1.

tions. We find that the results are robust when we choose items within the median difficulty level range.

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

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

[34] See Appendix O.

Figure 2: The Distribution of Denver Test Passed Items from Model with Item Difficulty Levels

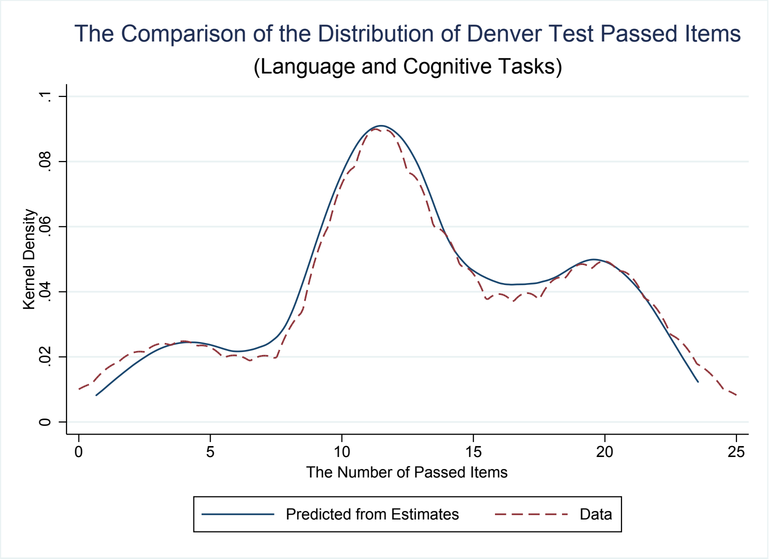
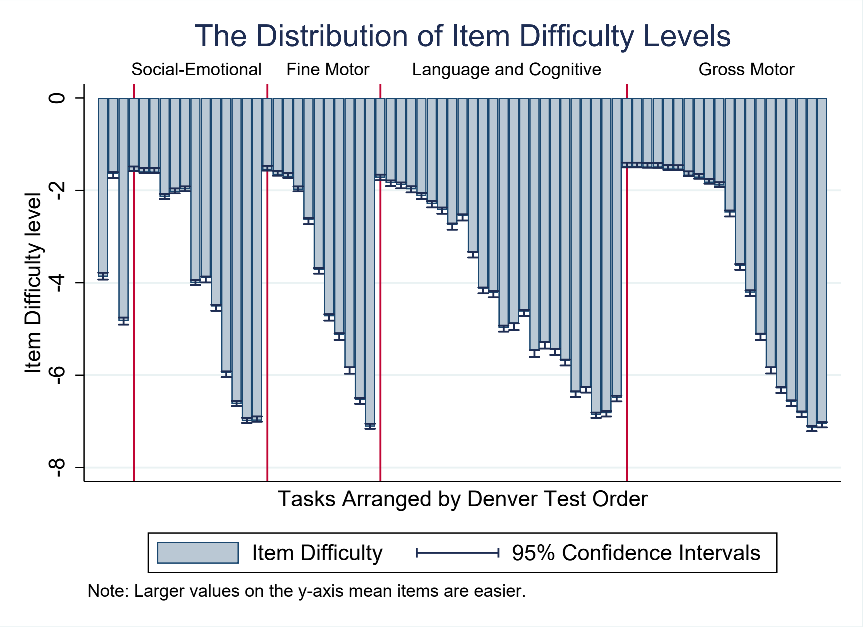


Figure 3 shows the array of estimated difficulty level parameters δ^{j_k} for each task item. When the item difficulty level increases, estimated scores become more negative. The estimates generally accord with the design of tests to increase the difficulty level with later items. The estimated difficulty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

An advantage of our approach is that we can estimate individual-level latent skill factors. First, Table 5 presents the treatment effects for the means of the four

Figure 3: The Distribution of Denver Task Item Difficulty Levels



Note: Larger values on the y-axis mean items are easier.

Table 5: 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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

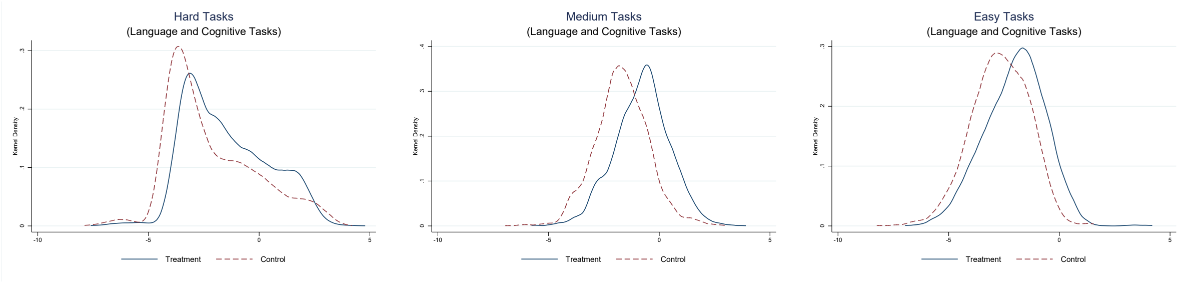
latent skill factors. Except for gross motor skills, the means of all other latent skill factors in the treatment group are statistically significantly larger than those in the control group. When we compare treatment effects across different latent skills, we find that improvements in fine motor and language skills are roughly the same but that there are no treatment effects for gross motor skills.[35]

Identifying factors and factor loadings is fraught with controversy regarding appropriate normalizations. Figure 4 plots effective skills—the product of estimated skill factor loadings and the latent skill factors $\boldsymbol{\theta}'\boldsymbol{\alpha}$ based on the Denver task difficulty levels for language and cognitive skills.[36] Estimating this term does not require any normalization. On average, the loadings for the treatment group are larger for all tasks whatever their difficulty, but the shift for the loadings of easy tasks is less clear. Figures Q.4–Q.6 in Appendix Q present the comparison of the distribution of $\boldsymbol{\theta}'\boldsymbol{\alpha}$ for treatment and control groups for other skills. The same pattern emerges. Effective skills are increased for the treated regardless of any normalization.

[35] Our Anderson-Rubin normalization assumes that latent skills are independent. Using alternative normalizations, we can identify the joint distributions of latent skills. See Carneiro et al. (2003) or Williams (2020).

[36] Figures Q.1 and Q.2 present the latent skill loadings on other types of tasks. Since we have 72 tasks in total, the tasks with the top 24 difficulty parameters are defined as easy tasks, the bottom 24 are defined as hard tasks, and the middle 24 are defined as medium tasks. All rankings are based on the estimates of the task difficulty level parameters.

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



(a) Hard Tasks

(b) Medium Tasks

(c) Easy Tasks

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

When we impose the Anderson-Rubin normalization, we generally reject the hypothesis that on average factor loadings are the same across treatment and control groups.[37] Table 6 reports tests of equality of the average loadings on different tasks for the different skills. Except for gross motor skills, we reject the hypothesis. The loadings on latent language and cognitive skills are large, but the loadings for social-emotional skills are smaller, suggesting that, on average, the program reduces the effectiveness of such skills.

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

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

Table 6: Estimated 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.	test of equality of mean loadings
Language and Cognitive	0.453	0.364	Language and Cognitive	0.679	0.469	0.000
Social-Emotional	0.259	0.263	Social-Emotional	0.222	0.246	0.002
Fine Motor	0.448	0.251	Fine Motor	0.556	0.211	0.001
Gross Motor	0.739	0.405	Gross Motor	0.693	0.442	0.276

Notes: 1. These are the means and standard deviations of $\boldsymbol{\alpha}^{j_k,\mathbf{0}}$ and $\boldsymbol{\alpha}^{j_k,\mathbf{1}}$, respectively, across items.

2. p-values are for the null of equality of treatment and control summary measures.

3.3.3 Comparisons with a Model without Task Difficulty Parameters

To show the impact of introducing task difficulty parameters into the model, we estimate a restricted version of the model based on Equation (3), in which we set all task difficulty parameters equal to zero. First, we compare the likelihood ratio between the full model and the restricted model and find that the full model has a higher likelihood. The likelihood ratio test statistic is $\chi^2(71) = 8419.26$. The p-value of rejecting the null hypothesis of equal task difficulty across items is less than 0.001. Figure 5 shows the worsening of the fit to sample test scores when difficulty levels are suppressed. We compare the restricted and unrestricted fits for other skills in Appendix O.

Second, we compare the treatment effects on the mean of latent skill factors in Table 7 ($E(\boldsymbol{\theta}^1) - E(\boldsymbol{\theta}^0)$). Notice that estimates of the model without task difficulty parameters are very different from the estimates with the difficulty parameters. A model without difficulty parameters produces significantly negative effects on social-emotional skills and significantly positive effects on gross motor skills, which are inconsistent with both the full model and the OLS model treatment effect evaluations. Adjusting for difficulty level matters.

Figure 5: The Distribution of Denver Test Passed Items from Model without Item Difficulty Levels

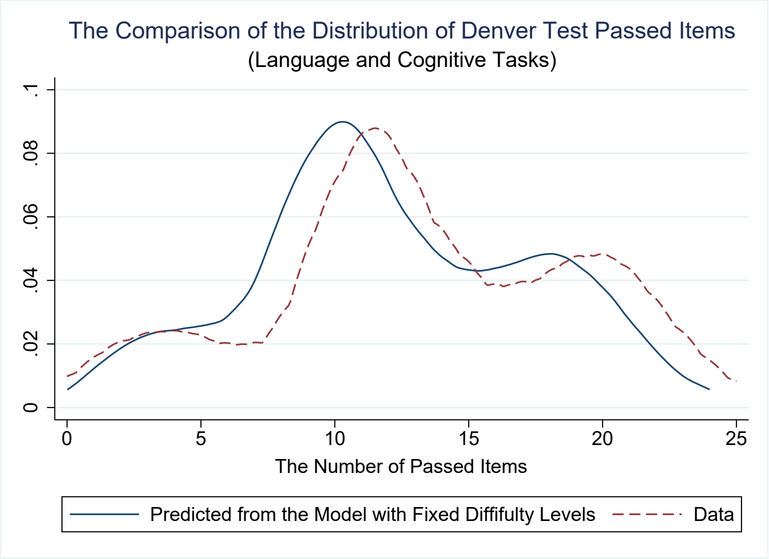


Table 7: Comparing Treatment Effects for $\boldsymbol{\theta}$ Based on Two Models with and without Difficulty Parameters

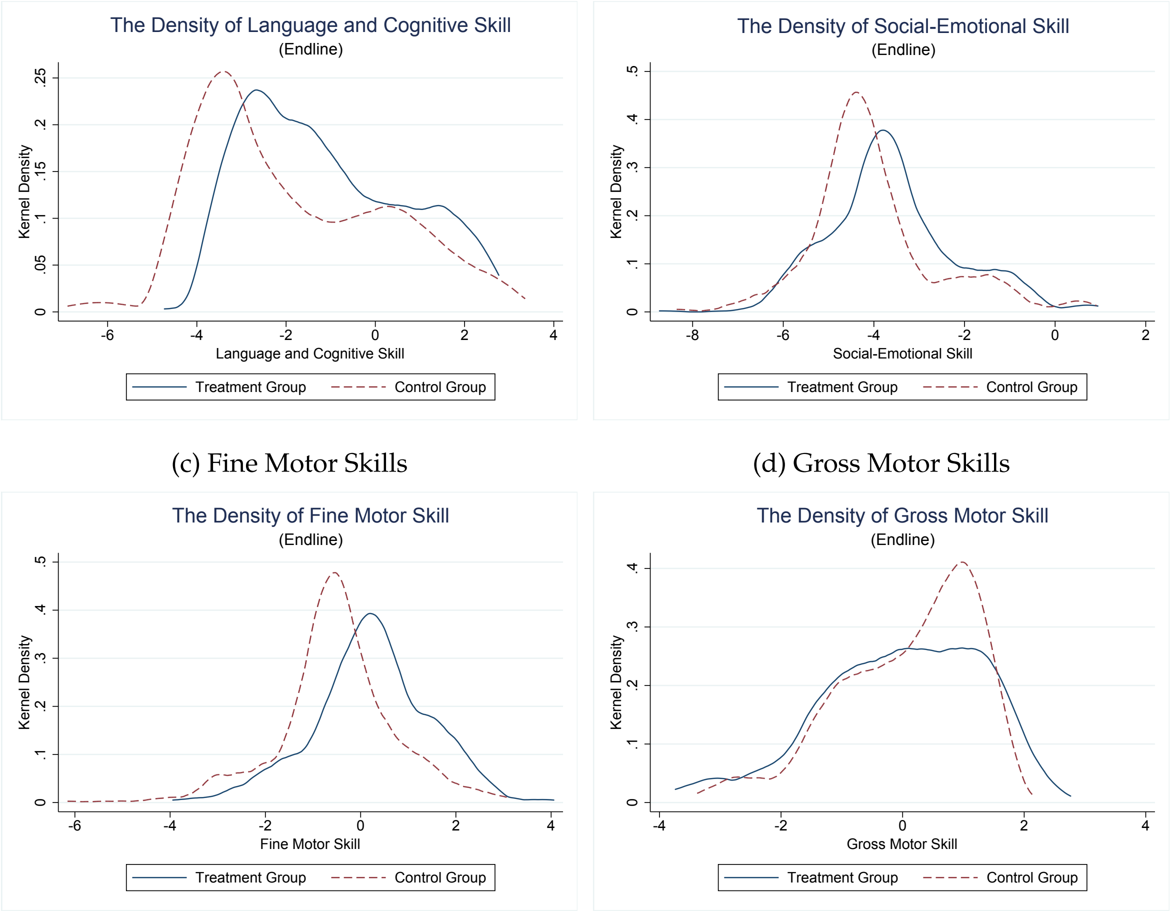
	Social-Emotional	Fine Motor	Language	Gross Motor
Full Model	0.395***	0.726***	0.753***	-0.095
(With Task Difficulty Adjustment)	[0.208, 0.583]	[0.551, 0.899]	[0.459, 1.051]	[-0.280, 0.089]
Restricted Model	-3.14***	1.136***	1.158***	1.069***
(Without Task Difficulty Adjustment)	[-3.375, -2.904]	[1.205,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

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

3.3.4 Distributions of Latent Skill

We next show that the intervention improves all but motor skills. We discover where in the baseline no treatment distribution improvement is greatest. We compare densities of skills of treatment and control groups at endline. Distribution functions are plotted in Appendix R. Figure 6a shows that the density of language and cognitive skills for the treatment group shifts right and has a fatter right tail than the one for the control group. Latent language and cognitive skill distributions are more right-shifted for the treatment group. Differences are more substantial at the bottom and middle of the treatment distribution compared to those at the top.

Figure 6: Treated and Untreated Skill Distribution


(a) Language and Cognitive Skills

(b) Social-Emotional Skills

(c) Fine Motor Skills

(d) Gross Motor Skills

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

For gross motor skills, there is little evidence of any treatment effect. The factor distributions are similar between the treatment and control groups. Figure 6d shows that the densities of the gross motor skills are very close to each other for treatments and controls.

In summary, language and cognitive, social-emotional, and fine motor skills were substantially improved by the program. Gains are not uniform across the control distribution for language and cognitive skills. They are roughly uniform for 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.

The patterns at midline are similar but growth is greater at endline. See Appendix R. Appendix S presents an extensive array of stochastic dominance tests for the estimated distributions, which produce have the same qualitative favor.

4 Decomposing ATE

We use our estimates of latent skill profiles to analyze the sources of the experimental ATEs. We compare experimental treatment effects with those obtained from our model. Average treatment effects produced by the experiment can arise either from changes in the mapping from skills to task performance or from changes in skills. We investigate the quantitative importance of each of these sources. Before doing so, we assess the performance of our skill estimates in predicting experimental treatment effects.

The latent outcome for skill j is:

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

Since we recover the individual latent skills $\boldsymbol{\theta_i^d}$, we can use them as inputs into our estimates of Equation (3) to simulate average treatment effects on Denver test scores in order to gauge the quality of our estimates. The point estimates of the average treatment effects so obtained are in close agreement (see Table 8).

Table 8: Average Treatment Effect Point Estimates Comparison

Denver Tasks	From OLS Model	From Factor Model	p-value of equality of estimates from the two models
	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
$\chi^2(4) = 0.116$			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.

4.1 The Sources of Our Treatment Effects

Experimental treatment effects may arise not only from enhancements of latent skills $\boldsymbol{\theta_i^d}$ but also from changes in the mapping from skills to task performance $\boldsymbol{\alpha}^{\boldsymbol{j_k,d}}$ and $\boldsymbol{\beta}^{\boldsymbol{j_k,d}}$. We examine whether such shifts explain a quantitatively important portion of estimated treatment effects. To do so, we decompose the item-level treatment effects into two components: the effects from the changes in the mapping from skills to tasks and the effects of treatment on skills.

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

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

where we assume $\varepsilon_i^{j_k} \sim N(0,1)$. Home visiting treatment effects arise from three channels: changes in the coefficient on observables $\boldsymbol{\beta}^{\boldsymbol{j_k,d}}$, changes in latent skill factors $(\boldsymbol{\theta_i^d})$, and changes in factor loadings for skills, the $\alpha^{j_{k,d}}$. 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:

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

Notice that Equation (6) holds over a common support for $\boldsymbol{X}$ and when the factors in the treatment and control groups have similar distributions of observable covariates, which conditions are essentially satisfied in our sample.[38] Table 9 reports the decomposition of treatment effects. The main drivers of the treatment effects are increases in latent skills. We previously noted that there is no significant difference in $\boldsymbol{\beta}$ between the treatment and control groups. The contribution to the treatment effects from $\boldsymbol{\beta}$ is accordingly negligible. We decompose the treatment effects in the order suggested in Equation (6). The contribution from experimentally induced changes in $\boldsymbol{\alpha}$ is not precisely estimated, despite the statistically significant shift in the $\boldsymbol{\alpha}$s documented in Table 6. For this reason, we conclude that the dominant effect of treatment is on latent skills. Section U in the appendix shows that decompositions conducted in different orders for different sets of family conditioning variables, produce similar qualitative and quantitative results.

Table 9: Source of Treatment Effects (Decompose Observed Covariates First)

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

Notes: 1. Total treatment effects for skill k are $\frac{1}{N_{J_k}}\sum_{j_k=1}^{N_{J_k}}\left(\frac{\sum_{i=1}^{N_I} Y^{j_k,i}D_i}{\sum_{i=1}^{N_I} D_i} - \frac{\sum_{i=1}^{N_I} Y^{j_k,i}(1-D_i)}{\sum_{i=1}^{N_I}(1-D_i)}\right)$ assuming both denominators are nonzero and N_I is the number of observations.
2. To ensure that the observed covariates are balanced between the treatment and control groups, we consider the sample of children who are younger than 46 months and older than 12 months.
3. Standard errors are reported in parentheses.

[38] 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 T, we show the age distribution between treatment and control groups.

4.2 Treatment Effects on Latent Skills Conditional on Caregiver Status

This section compares treatment effects based on the caregiver's characteristics. 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 10 provides treatment effects on latent skill factors $\boldsymbol{\theta_i}$. The table reveals that, at the endline, the largest treatment effects are for vulnerable children for whom mothers are absent (i.e., mother works outside or both parents work outside). In most cases, grandmothers with low levels of education are the caregivers when mothers are absent.[39] This result is similar to the findings of Bernal and Keane (2011) that out of home daycare is worse for cognition for the development of children except when home daycare is provided by grandmothers.

[39] See Appendix B.3.

Table 10: Treatment Effects on Latent Skills $\boldsymbol{\theta_i}$

	(1)	(2)	(3)	(4)			
Standardized	Non-Left-Behind Children	Left-Behind Children			Test Equality (p-value)		
		Mother Works Outside	Father Works Outside	Both Work Outside	(1)=(2)	(1)=(3)	(1)=(4)
			Midline				
Language and Cognitive	0.503***	0.730**	0.308*	0.671*	0.402	0.507	0.718
	[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	0.786	0.355	0.673
	[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	0.374	0.252	0.550
	[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	0.333	0.921	0.457
	[-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**	0.848	0.047	0.809
	[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***	0.026	0.180	0.000
	[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	0.035	0.195	0.000
	[-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	0.056	0.006	0.006
	[-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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
4. The columns of test equality present the p-values of the null hypothesis to test whether the treatment effect sizes are equal to that for non-left-behind children.

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

Table 11 shows that for comparable outcome measures at early ages, China REACH is on track with Jamaica Reach Up and Learn, which has been shown to generate substantial lifetime benefits (Gertler et al., 2014; Grantham-McGregor and Smith, 2016). We cannot reject the hypothesis that the treatment effects are the same across these two interventions, except for motor skills. Recall that Jamaica targeted a stunted population. If China REACH continues on course, it should reproduce the effects of the successful Jamaica program. See the analysis in Zhou et al. (2023).

Table 11: Treatment Effects on China REACH and Jamaica Reach Up and Learn

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

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

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

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

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

6 Contributions of this Paper to the Literature on Reach Up

There is a large and growing literature on applications of Jamaica Reach Up to a variety of less developed countries. Jervis et al. (2023) presents a meta-analysis of this literature. Meta-analytic studies have important limitations. However, the Jervis analysis is much more convincing than most. Measures used are comparable across studies. Appendix V reports the unadjusted effect sizes of the meta-analytic studies. The unadjusted standardized treatment effects we report in the first part of the paper are well within the range of those reported in her study, except we find little impact of the program on the home environment, whereas other studies report stronger effects.

There are no studies like ours that estimate impacts on distribution of latent skills. We demonstrate the large improvement in fit that arises from allowing latent skills to affect measured test outcomes. None of the studies accounts for item difficulty which we have shown is important. Doing so reverses some of the findings for the type of unadjusted outcomes reported in the Jervis et al. (2023). survey. None of the papers in her survey studies the mechanisms through which the program operates. We show that it largely operates through enhancing skills rather than in enhancing use of existing skills.

The generally positive results reported by Jervis et al. (2023), in conjunction with the evidence in this paper strongly supports a beneficial effect of Reach Up on its participants, at least in the short run. This paper points to better ways to analyze and interpret the existing data than have been used in previous studies.

7 Conclusion

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

Our analysis offers a prototype for measuring latent skills using diverse outcome measures that adjust for the difficulty inherent in different items (tasks). Our adjustments produce more plausible estimates. We estimate child latent skills and how they are affected by the program. We develop a framework for understanding the mechanisms generating treatment effects. We test and reject the "dedicated factor" measurement model widely used in the economics of skill formation (e.g., Agostinelli and Matthew (2023)). Measured item scores depend on multiple skills.

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

References

Agostinelli, F. and W. Matthew (2023). Estimating the technology of childrens skill formation. Forthcoming, *Journal of Political Economy*.

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.

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *Conditionally accepted by the American Economic Review 112*(12), 3911–3940.

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

Bernal, R. and M. P. Keane (2011, July). Child care choices and children's cognitive achievement: The case of single mothers. *Journal of Labor Economics 29*(3), 459–512.

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.

Cameron, S. V. and J. J. Heckman (1998, April). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males. *Journal of Political Economy 106*(2), 262–333.

Cameron, S. V. and J. J. Heckman (2001, June). The dynamics of educational attainment for black, Hispanic, and white males. *Journal of Political Economy 109*(3), 455–499.

Canay, I. A., A. Santos, and A. M. Shaikh (2021). The wild bootstrap with a “small” number of “large” clusters. *Review of Economics and Statistics*, 1–45.

Carneiro, P., K. Hansen, and J. J. Heckman (2003, May). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review 44*(2), 361–422.

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. and J. J. Heckman (2008, Fall). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources 43*(4), 738–782.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010, May). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Elango, S., J. L. García, J. J. Heckman, and A. Hojman (2016). Early childhood education. In R. A. Moffitt (Ed.), *Economics of Means-Tested Transfer Programs in the United States*, Volume 2, Chapter 4, pp. 235–297. Chicago: University of Chicago Press.

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

Gertler, P., J. J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. M.

Chang, and S. Grantham-McGregor (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. *Science 344*(6187), 998–1001.

Grantham-McGregor, S. and J. A. Smith (2016). Extending the jamaican early childhood development intervention. *Journal of Applied Research on Children: Informing Policy for Children at Risk 7*(2).

Heckman, J. and J. Zhou (2022a). Interactions as investments: The microdynamics and measurement of early childhood learning. Under revision, *Journal of Political Economy*.

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

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

Heckman, J. J. (1981). Statistical models for discrete panel data. In C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 114–178. Cambridge, MA: MIT Press.

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

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

Howard, K. S. and J. Brooks-Gunn (2009). The role of home-visiting programs in preventing child abuse and neglect. *The Future of Children 19*(2), 119–146.

Jervis, P., J. Coore-Hall, H. O. Pitchik, C. D. Arnold, S. Grantham-McGregor, M. Rubio-Codina, H. Baker-Henningham, L. C. Fernald, J. Hamadani, J. A. Smith, and Others (2023). The Reach Up parenting program, child development, and maternal depression: A meta-analysis. *Pediatrics 151*(Supplement 2).

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

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

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika 49*, 115–132.

Rabe-Hesketh, S. and A. Skrondal (2016). Generalized linear latent and mixed modeling. In W. J. van der Linden and R. Hambleton (Eds.), *Handbook of Item Response Theory: Models, Statistical Tools, and Applications*, Volume 1, Chapter 30, pp. 531–554. Boca Raton, FL: Chapman and Hall/CRC.

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.

von Davier, M. (2016). Rasch model IRT. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory*, Volume 1, pp. 31–48. Boca Raton, FL: Chapman and Hall/CRC.

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

Welch, F. (1970). Education in production. *78*(1), 35–59.

Williams, B. (2020). Identification of the linear factor model. *Econometric Reviews 39*(1), 92–109.

Zhou, J., J. J. Heckman, B. Liu, M. Lu, S. M. Chang, and S. Grantham-McGregor (2023). Comparing china REACH and the Jamaica home visiting program. *Pediatrics 151*(Supplement 2).